COMPUTATIONAL STUDIES OF STRUCTURE-FUNCTION RELATIONSHIPS OF FAMILY 48 CELLULASES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirement for the Degree of

Doctor of Philosophy

by

Mo Chen

January 2014

COMPUTATIONAL STUDIES OF STRUCTURE-FUNCTION RELATIONSHIPS OF FAMILY 48 CELLULASES

Mo Chen, Ph.D.

Cornell University 2014

ABSTRACT

Cellulases are key enzymes in lignocellulosic biomass degradation for producing ethanol as biofuel. However, the process of biomass conversion through the approach of enzyme treatments is notoriously inefficient. In particular, biomass saccharification by cellulases is the rate-limiting step. This is partially due to the low turnover number of cellulases. This research was focused on attempting to find a fundamental understanding of the structure-function relationships of cellulases, and specifically family 48 cellulases, which are a major group of processive exocellulases capable of hydrolyzing crystalline cellulose.

The characterization of multiple crystallographic structures of family 48 cellulases has revealed that their structures have two interesting features. One is that the six inner α-helices of a $(\alpha/\alpha)_6$ barrel structure form a water-filled pore that connects the active site with the protein surface, and the other is a Trp-rich active site tunnel that accommodates a cellooligomer substrate chain which takes up the substrate binding subsites from -7 to +2, indicating that they produce mostly a cellobiose product at each catalytic cycle.

With regard to the water pore structure, it was hypothesized that this pore structure might be of mechanistic importance in transporting water molecules for substrate hydrolysis. Molecular dynamics simulations were used to study the water flow in the pore, and in theoretical mutants, particularly those with mutations at the sites of conserved residues and at the inner sections of the

pore converted into Phe's, which were designed to block the water passing through the pore for experimental test on their enzymatic activity. Unfortunately, the mutants were not folded properly experimentally, thus failing to test the hypothesis.

Product inhibition in cellulases has been observed in experimental studies and molecular simulations, and thus potentially contributes to the low activity of these enzymes. The mechanism of product inhibition might be that the product, mainly cellobiose, binds to the active site tunnel exit, preventing the enzymes from progressing along the cellulose chain for further hydrolysis. The binding affinity for cellobiose to the cellulase tunnel exit was investigated. Rational mutants of the cellulases with the mutation sites at the tunnel exit were designed and evaluated, aiming at reducing the product inhibitory level.

Imidazole, the sidechain of histidine, exists frequently in carbohydrate binding proteins, including some cellulases, and was hypothesized to interact with the glucose repeat unit of cellulose in a particular fashion. The interaction of imidazole with glucose in aqueous solution was investigated using MD simulations, and it was found that imidazole form both stacking interactions and hydrogen-bonding interactions with the beta anomer of D-glucose.

BIOGRAPHICAL SKETCH

Mo Chen earned her Bachelor of Engineering degree in Food Science and Engineering from Beijing Forestry University in 2006. She received her Master of Science degree in Food Science in 2008 from University of Missouri at Columbia, Missouri. In 2008, she joined the doctoral program at Brady Lab in Department of Food Science at Cornell University.

Mo Chen's doctoral research is focused on computational atomic modeling and simulations of biomolecules for bioethanol production, which is in collaboration with the National Renewable Energy Lab (NREL) as part of the BioEnergy Science Center (BESC) project. She has presented her research at the BESC Science Retreats. Part of her research has been published in *Carbohydrate Research*, and the rest has been or will be submitted to other scientific journals.

Mo Chen's dissertation, *Computational Studies of Structure-Function Relationships of Family 48 Cellulases*, was supervised by Dr. John W. Brady.

This document is dedicated to Chen Jinlian and Li Yuhua.

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my committee chair, Professor John Brady, who is a very wise and knowledgeable scientist, and has the attitude of caring about science and nature comprehensively. Without his guidance and persistent help, this dissertation would not have been possible.

I would like to express my thanks to my committee members, Professor David Wilson and Professor Robert Oswald, for their academic advice and guidance during my research and study. In particular, the meetings, collaborations, and discussions on cellulases with Professor David Wilson and his group have significantly enlightened my work on the computational studies of cellulases.

I would like to thank Dr. Yannick Bomble, Dr. Michael Crowley, and Dr. Michael Himmel from the National Renewable Energy Lab for their inputs in the collaboration on the cellulase projects. And I want to thank the National Renewable Energy Lab for providing funding and computer resources for my research.

I would like to thank Dr. Maxim Kostylev, Dr. Phil Mason, Dr. Jakob Wohlert, Dr. Malin Wohlert, Dr. Peter Hansen, Dr. Udo Schnupf, Dr. Linghao Zhong, Dr. James Matthews, and other previous lab members for the opportunities of collaborations and for their help with my research.

My appreciation is also extended to all my friends at Cornell for all their friendship over the past five years.

Finally, I would also like to express my thankfulness to my boyfriend, Dr. Xavier Serey, who has been very supportive to me and has helped me significantly in my study. And I am forever thankful for the love and support of my parents, and my aunt Jinchen Chen.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1. Biofuel production and lignocellulosic biomass

Lignocellulosic biomass exists abundantly in nature, in the form of dedicated energy crops (such as switchgrass, big bluestem and Indian grass), forestry residues, and agricultural residues. It is one of the most promising renewable resources for the production of biofuel. Lignocellulosic biomass can be processed into biofuels in three ways: 1) hydrolysis to sugar and subsequent fermentation, 2) gasification to synthesis gas and its subsequent conversion through the Fischer-Tropsch process or fermentation, and 3) pyrolysis to char, bio-crude or gas [1]. The biofuel products from lignocellulose include gasoline, diesel fuel, jet fuel, and ethanol, among which ethanol is the main product so far.

The degradation of lignocellulosic biomass through enzymatic treatments is considered the most promising method of bioethanol production, and has been studied extensively. In principle, biomass degradation involves three steps: pretreatment that separates and removes the lignin and hemicellulose components; enzymatic treatment that hydrolyzes cellulose into soluble cellodextrins and glucoses; and fermentation that converts glucose into liquid biofuels. The biomass conversion process faces many difficulties in being scaled up because of its inefficiency due to biomass recalcitrance and slow enzyme activities.

Lignocellulose is composed of heterogeneous intertwined cellulose chains, hemicellulose, and lignin. Cellulose is the major component of lignocellulose. It is a linear polysaccharide polymer consisting of β-(1→4) linked D-anhydroglucopyranose, and its repeating unit is anhydrocellobiose. The degree of polymerization (DP) of cellulose chains ranges from 100 to 20,000 [2]. Native cellulose, referred to as cellulose I, has two distinct crystalline forms, Iα and Iβ. Iα is dominant in bacterial and algal cellulose, while Iβ is dominant in higher plants [3]. Native cellulose can be converted to other crystalline polymorphs (cellulose II – IV) through various treatments [4].

Hemicellulose is a branched polysaccharide characterized by β-(1→4) linked pentoses and hexoses, such as glucose, xylose, mannose, galactose, and arabinose. Lignin is a complex compound that lacks a defined primary structure. It is a heterogeneous mixture of dendritic polymers that are composed of hydroxyphenyl propanoids [5, 6]. In lignocellulose, cellulose is partially crystalline and cellulose chains bundle together in a parallel manner in an oriented pattern forming elementary fibrils. Brunches of the elementary fibrils are embedded in a matrix of hemicellulose, which glue the cellulose elementary fibrils together, thus enhancing the strength of plant cell walls [7]. Lignin is located on the exterior of the matrix covalently bonded to hemicellulose, and is resistant to biochemical conversion. Collectively these polymers make up structural units known as microfibrils, whose dimension ranges from 10-20 nm in diameter [8].

Cellulose can be hydrolyzed to glucose by enzymatic treatment, and then fermented into bioethanol as one important form of renewable biofuel. It is a crucial task for ethanol production to hydrolyze cellulose into glucose in a cost- and energy-efficient manner. The highly crystalline nature of cellulose hinders its degradation. The crystallinity of native cellulose can be described by a crystallinity index (CrI). The most common and simple method to determine CrI is by using the X-ray diffraction peak height method [9]:

$$CrI = 100 \cdot \frac{I_{200} - I_{non-cr}}{I_{200}} \ [\%]$$

where I_{200} refers to the maxim intensity of the crystalline peak corresponding to the plane in the sample with the Miller indices 200 and I_{non-cr} refers to the intensity the diffraction of the non-crystalline cellulose . Enzymatic hydrolysis for cellulose with lower CrI is typically much faster than for cellulose with higher CrI [10]. Computational modeling of cellulose has revealed that both the hydrophobic association between cellulose chains and hydrogen bonding favor the crystal packing [11]. Certain acids and enzymes can depolymerize cellulose, both with pros and cons. For example, concentrated sulfuric acid (>65%) is able to effectively hydrolyze crystalline cellulose at moderate temperatures, but causes equipment corrosion [12]. Diluted acid (0.5-1%) can be used for cellulose

hydrolysis, but requires high temperatures (150-220°C) and leads to the production of undesirable side products that inhibit fermentation of glucose to ethanol [13]. The cellulolytic enzymes, including cellulases and β-glucosidase, hydrolyze cellulose into glucose at much lower temperatures (40-85°C). In general, enzymatic treatment is commonly accepted in cellulosic biomass conversion for biofuel production, though it is relatively costly as the reaction rate of cellulose hydrolysis is very low and bulk usage of cellulases is needed. The cost of cellulase was most recently estimated to be \$0.46 gal^{-1} (\$0.12 L^{-1}) of ethanol in a 150 million liters per year (MMLY) plant using 11.5 mg enzyme g^{-1} substrate [14]. It is important to reduce both the cost and the amount of cellulases used. Recycling of cellulases provides an option to reduce enzyme loading [15]. Designing cellulase cocktails with improved efficiency and modifying cellulases through mutagenesis also serve as strategies.

The crosslinking between the polysaccharides and lignin via ester and ether linkages also gives rise to recalcitrance of lignocellulose to biodegradation, as the existence of hemicellulose and lignin in the microfibrils reduces the accessibility of cellulose to hydrolytic enzymes. Genetic modifications have been used to steer biosynthetic pathways toward designing plants that either grow less lignin or produce lignin that is more susceptible to chemical degradation, though fitness issues of such plants potentially exist [16]. Several pretreatment techniques have been developed to remove lignin and hemicellulose in lignocellulose, including stream explosion pretreatment [17], dilute acid pretreatment, and ammonia recycle percolation [18]. In particular, a novel pretreatment approach named COSLIF (cellulose solvent- and organic solvent-based lignocellulose fractionation) was able to fractionate lignocellulose into cellulose, hemicellulose, and lignin under modest reaction conditions (50°C and atmospheric pressure), and the highest digestibility of the pretreated biomass reached ~97% in 24 h at the enzyme loading of 15 filter paper units of cellulase and 60 IU of β-glucosidase per gram of glucan [19]. Recent advances in ammonia pretreatment using liquid ammonia resulted in production of a cellulose III polymorph, which significantly enhanced the

effectiveness of enzymatic depolymerization [20, 21]. Many ionic liquids are able to dissolve lignocellulose, and the cellulose component can be regenerated by the addition of water or anti-solvents, whereas the majority of lignin and hemicellulose remains dissolved in the ionic liquids and can be recovered afterwards [22, 23]. Enzymatic hydrolysis of the regenerated cellulose from ionic liquids gives high yields within significantly less time than that required for dilute acid pretreated biomass [24]. However, the cost of ionic liquids is high due to the challenges faced in the synthesis. Additionally, alkaline pretreatment using sodium hydroxide or calcium hydroxide is effective in removing lignin from the biomass. For example, cotton stalks pretreated with sodium hydroxide was reported to have high level of delignification of (65.63% for 2% NaOH, 90 min, 121 °C/15 psi) [25]. Coastal Bermuda grass pretreated with 0.5~3% NaOH resulted in up to 86% lignin removal [26]. Ultrasound treatment can induce cavitation in biomass and therefore increases the biomass surface area, enhancing its accessibility to hydrolytic enzymes. It was reported that ultrasound intensively enhanced cellulose dissolution in ionic liquids, to the level of complete dissolution within a few minutes [27].

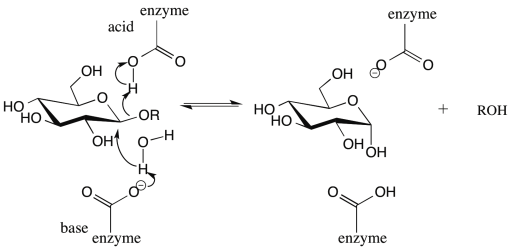
1.2. Enzymes for lignocellulose degradation

1. 2.1. Enzyme classifications

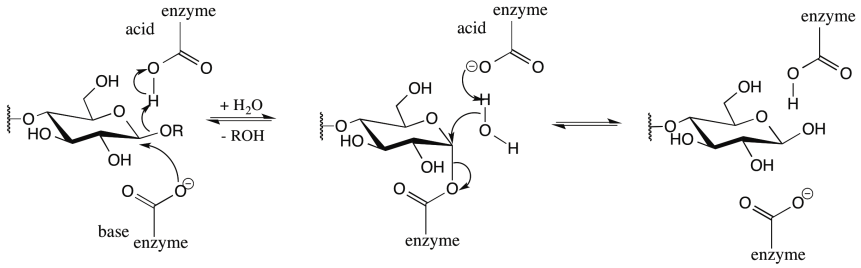
Cellulases are a type of glycosyl hydrolase. Based on the mode of catalytic reaction and on the substrate specificity, glycosyl hydrolases are categorized as (EC 3.2.1.x) by the Enzyme Nomenclature established by the International Union of Biochemistry and Molecular Biology (IUBMB). Here "3.2.1" refers to the enzymes that hydrolyze *O*-glycosyl linkages, and the last number designates the substrate and sometimes the molecular mechanism. Glycosyl hydrolases follow two mechanisms, resulting in inversion or retention of the stereochemistry at the product cleavage site (Figure 1.1) [28]. The glycosyl hydrolases involved in cellulosic biomass degradation for biofuel production mainly include:

- Endoglucanases or 1,4-β-D-glucan-4-glucanohydrolases (EC 3.2.1.4)
- Cellobiohydrolases or 1,4-β-D-glucan cellobiohydrolases (EC 3.2.1.91, EC 3.2.1.176)
- β-glucosidases or β-glucoside glucohydrolases (EC 3.2.1.21)

Endoglucanases hydrolyze the glycosidic bonds of accessible cellulose chains randomly to produce shorter cellulose chains. As a result, their activity increases the viscosity of the substrate solution. Cellobiohydrolases, also known as exocellulases, are a kind of exoglucanases. They hydrolyze cellulose chains either from their reducing ends or nonreducing ends, producing predominantly cellobiose and a perceptible amount of cellotriose and cellotetraose. Endocellulases and cellobiohydrolases can disrupt the crystalline cellulose surface, making it vulnerable to enzymatic degradation. In particular, cellobiohydrolase activity is essential to hydrolysis of microcrystalline cellulose [29]. Overall, the course of cellulose depolymerization induced by endocellulases and cellobiohydrolases is typically the rate-limiting step in the process of cellulosic ethanol production. β-glucosidases can efficiently hydrolyze cellobiose and other soluble cellodextrins (DP≤6) into D-glucose. In addition, another kind of exoglucanase named 1,4-β-D-glucan glucanohydrolases (cellodextrinases) (EC 3.2.1.74) has been found in cellulolytic microbes, where they produce glucose as the main product of cellulose hydrolysis.



(a) Inverting mechanism



(b) Retaining mechanism

Figure 1.1. The two mechanisms of glycosyl hydrolases.

Another classification method developed by Henrissat and coworkers categorizes glycosyl hydrolases into different families based on protein amino acid sequence similarities [30-32]. Because a direct relationship between protein sequences and folding similarities is well recognized [33], this method is advantageous in that it provides a way to identify 3 dimensional structural similarities of protein molecules. In this method, the families with similar 3 dimensional structures are grouped into “clans”, indicating evolutionary aspects of carbohydrate metabolism [30-32]. Glycosyl hydrolases of the same family appear to have the same reaction mechanism [34].

1.2.2. Cellulases

Cellulolytic microorganisms include a range of anaerobic bacteria, aerobic bacteria, and fungi. The cellulolytic aerobic bacteria and fungi have high cell yields and produce substantial amounts of noncomplexed cellulase systems for cellulose hydrolysis, whereas the cellulolytic anaerobic bacteria primarily produce complexed cellulosomal assemblies for cellulose degradation [29]. An up-to-date database for carbohydrate-active enzymes is available online at www.cazy.org.

1.2.2.1 Noncomplexed cellulases

Noncomplexed cellulase systems are referred to as a cellulase mixture in which each component can function discretely. They are composed of endoglucanases and cellobiohydrolases. Among the microorganisms that produce noncomplexed cellulase systems, significant attention has been drawn to the aerobic fungal *Trichoderma* species. In particular, *T. reesei* was found to produce at least two exoglucanases (CBHI and CBHII), and five endoglucanases (EGI, EGII, EGIII, EGIV, and EGV) [35, 36]. CBHI and CBHII cleave cellulose chains from their reducing ends and nonreducing ends, respectively. CBHI, CBHII, and EGI are each composed of a catalytic domain (CD) and a cellulose-binding domain (CBM), and a glycosylated polypeptide linker that connects the two domains. CBHI and CBHII feature a tunnel-shaped topology, serving as a substrate pathway during the processive action [37-39], whereas the active site of EGI is a groove instead of a tunnel [32]. CBHI and EGI have a retaining mechanism, whereas CBHII has an inverting mechanism. Due to the highly homologous similarity between CBHI and EGI, they are classified into family 7 glycosyl hydrolases [30]. The presence of a CBM is only essential for the cellobiohydrolase activity [40]. In addition, some other fungi such as *Humicola insolens* [41, 42] and *Phanerochaete chrysosporium* [43, 44], and some aerobic bacteria including the genera *Cellulomonas* [45] and *Thermobifida* [46, 47], have been thoroughly studied for their production of noncomplexed cellulase systems.

1.2.2.2. Cellulosomes

Cellulosomes are multi-enzyme complexes, and can be described as one of nature's most elegant and highly efficient supermolecular complexes. Anaerobic bacteria are the major microbial producers of cellulosomes. Some anaerobic fungi can also produce cellulosomes [48, 49]. The first cellulosome was found in *Clostridium thermocellum* in the early 1980s by Bayer and Lamed, and their coworkers [50, 51]. Other bacteria that produce cellulosomes include *Clostridia* species such

as *C. cellulolyticum* [52, 53], *C. cellulovorans* [54], and some *Ruminococcus* species such as *R. flavefaciens* [55-57]. Anaerobic fungi of certain genera such as *Neocallimastix* and *Piromyces* are reported to produce cellulosomes [58, 59]. Studies on cellulosomes have been comprehensively reviewed [60, 61].

The *C. thermocellum* cellulosome has a diameter of 18 nm and a mass in excess of 2 MDa [51, 62]. It consists of a large noncatalytic "scaffoldin" subunit CipA that carries nine type I "cohesion" (CohI) modules, a C-terminal type II "dockerin" (DocII) module, an "X module", and a cellulose-specific carbohydrate-binding module (CBM) [63, 64]. Each of the CohI modules mediates specific binding with a catalytic subunit through a complementary dockerin module that is possessed by the catalytic subunit. The catalytic subunits are composed of endocellulases and exocellulases. They hydrolyze cellulose into cellobiose and other soluble cellodextrins, which in the native environment are transported into *C. thermocellum* cells by ATP-binding cassette transports, and further hydrolyzed into glucose by intracellular phosphorylases for bacterial utilization [65]. The DocII module at the C-terminal end promotes the attachment of the cellulosome to the bacterial cell wall. The CBM assists in anchoring the cellulosome onto the surface of the crystalline cellulose substrate. This scaffoldin is the primary scaffoldin in *C. thermocellum*. In addition, *C. thermocellum* produces multiple anchoring scaffoldins, including four that contain type II cohesion domains (namely SdbA, OlpB, Orf2, and Cthe_736) and two that contain type I cohesion domains (namely OlpA and OlpC) [66-69]. These anchoring scaffoldins are located in the *C. thermocellum* cell wall, and mediate the attachment of the cellulosomes to the bacterial cell wall.

In general, the cellulosome complexes of *C. thermocellum* are secreted to the extracellular environment, where they mostly bind tightly to the bacterial cell wall while finding their way to attach to cellulose and hemicellulose, allowing for cellulolytic degradation in the proximity of the bacterium. The cellulosomes are released from the bacterium in the latter part of the growth phase. The *C. thermocellum* cellulosome was reported to display a higher specific activity on crystalline

cellulose than that of the noncomplexed *Trichoderma* system [70]. Mutations in the scaffoldin gene *CipA*, did not influence the hydrolysis activity of the cellulosome on soluble β-glucan, although did reduce the capacity of *C. thermocellum* to hydrolyze crystalline cellulose by 15-fold, supporting the importance of the cellulosomal complex in efficient hydrolysis of crystalline cellulose [71]. The high efficiency of the cellulosome might be attributed to the simultaneous catalytic events carried out by the adjacent cellulase subunits, as they promote synergistic effects on cellulose degradation.

The cohesion-dockerin binding events are mainly driven by hydrophobic interaction. The binding between cohesion and dockerin modules is species specific, in that type I dockerins do not interact with type II cohesions, and vise versa, allowing cellulosomes to assemble correctly and to attach to the bacterial cell surface in a certain manner [72]. The structure of the type I cohesion-dockerin complex in *C. thermocellum* reveals that the dockerin consists of three α-helices, in which helix 1 and helix 3 exhibit twofold symmetry with respect to amino acid sequence and a calcium-binding motif, and in which the cohesion module presents a flattened β-barrel structure [73]. Helix 1 and helix 3 of the dockerin contain an adjacent amino acid Ser-Thr pair, which plays a critical role in cohesion specificity and has a dual binding mode, as mutating the Ser-Thr pair into Ala's in helix 3 allows cohesion recognition to switch to helix 1 rather than helix 3 [74], and the docker-cohesion binding affinity is reduced only when the Ser-Thr pair is substituted in both helix 1 and helix 3 [75]. Calcium is of key importance in the stability of the type I dockerin. The type II cohesion-dockerin complex in *C. thermocellum* exhibits tighter binding than the type I cohesion-dockerin complex, and does not show a dual mode binding, though the dockerin has a twofold helical structure. Type II dockerin often occurs at the C-terminal end of the module named X-mode, which contributes to the stability of the dockerin [76].

1.2.3. Processivity of cellulases

The processivity of cellulases is defined to be the average number of cleavages that a cellulase carries out on a cellulose chain before it dissociates from the chain [77]. The processivity of cellulases contributes to the degradation of crystalline cellulose through detaching single cellulose chains from the surface and preventing them from reassociating with the solid part. Substituting the sugar-binding Trp's at the entrance of the active site tunnel in the exocellulase *T. fusca* Cel48A with Ala strongly decreased the cellulase activity [78]. However, improving the processivity of cellulases does not seem to be effective in increasing the enzymatic activity or the synergism between several cellulases [79].

Two types of processive cellulases have been identified, namely cellobiohydrolases and processive endoglucanases [77]. All the cellobiohydrolases seem to contain a hydrophobic tunnel structure in which the active site is buried. The tunnel leads the processive movement of a cellulose chain within it, and triggers its cleavage to produce a cellobiose from the chain end. The processivity of cellobiohydrolases can be estimated by measuring the ratio of cellobiose to cellotriose produced by the cellobiohydrolases [79]. This method is based on the assumption that for the first cleavage by an exocellulase there is an equal chance to produce either a cellobiose or a cellotriose, depending on the stereochemistry of the cellulose chain end. The processivity of *Clostridium phytofermentans* on Avicel (CrI 0.5~0.6) and regenerated amorphous cellulose was estimated to be about 3.5 and 6.0 [80]. The processivity of *Thermobifida fusca* exocellulase Cel6B was 7.2 on filter paper (CrI ~0.45) [79]. Several processive endoglucanases have been identified. One processive endoglucanase is found in *T. fusca* and contains a GH9 catalytic domain combined with a CBM domain, without which no processivity was detected [81, 82].

1.2.4. Activity of cellulases

In general, cellulase activity is measured using the IUPAC standard filter paper assay (FPA), which employs the dinitrosalicylic acid (DNS) method to determine the reducing sugars released

from 50 mg of Whatman #1 filter paper by cellulases [83]. This method has been further automated by Decker and coworkers [84]. The cellulase activity assay is used to screen cellulase mutants for performance.

Under optimal conditions, the activity of a cellulase also depends on the substrate. The activities of cellobiohydrolases on various cellulosic materials are all extremely low. For example, the activity of *T. fusca* Cel48A on the substrates swollen cellulose, CM-cellulose, BMCC, and filter paper were 0.40, 0.29, 0.19, and 0.07 μmol CB·min^{-1}·μmol enzyme^{-1}, respectively [85]. According to Kostylev and Wilson [86], the activity of *T. fusca* Cel48A on swollen cellulose can be estimated to be 1.16 μmol CB·min^{-1}·μmol enzyme^{-1}.

1.2.5. Product inhibition in cellulases

The activities of cellulases are competitively inhibited by the product cellobiose, causing a major obstacle for high product yields in hydrolysis of crystalline cellulose. Such inhibition has been observed in the exocellulase Cel7A from *T. reesei* [87-89] and family 48 exocellulases including *C. thermocellum* CelS [90-92] and *T. fusca* Cel48A [85]. Near-complete inhibition of the *C. thermocellum* cellulosome was observed at a low concentration of 2% (w:v) cellobiose [91]. It is found that the inhibition level in *T. reesei* Cel7A strongly depends on the nature of the substrate, whereas the endoglucanases from *T. reesei* are not affected by product inhibition [89]. Such phenomenon, to some extent, can be elucidated by the calculated product expulsion energies, which show that cellobiose tends to bind to the cellobiohydrolases more strongly than to the endoglucanases of *T. reesei* [93]. A study that has successfully reduced product inhibition by rational design involves the endoglucanases Cel5A from *Acidothermus cellulolyticus*. When converting the Tyr245, which is a residue that binds to the leaving group of the substrate (cellobiose), to Gly, there was significant decrease in product inhibition in that the K_I for the mutant was more than 1480% of that for the wildtype, and the hydrolytic activity was increased by 40%,

suggesting product inhibition affected enzymatic activity in a negative way [94].

1.2.6. Improvement on cellulases

Improvements of cellulases are desirable, with the emphasis on higher catalytic efficiency for insoluble cellulosic substrates, enhanced stability at elevated temperature and at certain pH values, and higher tolerance of product inhibition. The methods to modify single cellulases include rational design, directed evolution, and semi-rational protein engineering. In addition, models to promote synergism between several cellulases or the cellulase components of cellulosomal assemblies have been proposed.

Rational design was the earliest protein engineering approach, in which point mutants are made using the technique of site-directed mutagenesis. This strategy requires detailed knowledge of the three-dimensional protein structure, understanding of the protein's structure-function relationship, and a way to examine the characteristic changes of the mutants [95]. It aims at sorting out the direct correlation between the changed structures and changed functions. Rational design of cellulases often starts with mutating conserved non-catalytic protein residues near the active site, as these residues profoundly affect the enzymatic activity and substrate specificity [96]. One successful example was reported by Baker and coworkers [94], in which they mutated Tyr245 of the endocellulase Cel5A from *Acidothermus cellulolyticus* to Gly by rational design based on the understanding of the crystal structure, and achieved a 40% increase in the rate of cellulose hydrolysis. However, the majority of rationally designed cellulase mutants do not lead to significant improvement as expected.

Directed evolution has also been largely adopted in the field of protein engineering over the past two decades. It uses random mutagenesis (such as error-prone PCR) or recombinant DNA methods (such as DNA shuffling) to establish a library of mutated genes followed by library protein expression. Next, the performances of the protein mutants are evaluated by screening selections.

Directed evolution does not require a knowledge of the enzyme structure or its interaction with the substrates, though the screening methods are critical in selection of ideal mutants. Many studies have reported successful improvement of cellulases' properties using directed evolution. A *Bacillus subtilis* endoglucanase mutant generated by DNA shuffling was reported to display a five-fold higher specific activity [97]. A *T. reesei* EG III mutant generated by error-prone PCR exhibited an optimal pH of 5.4, which corresponded to a basic pH shift of 0.6 compared to the wildtype [98].

The approach of semi-rational protein engineering is an improvement to the method of directed evolution. Compared with directed evolution that establishes large combinatorial libraries before screening, the semi-rational protein design allows the selection of desired mutants through small, functionally rich libraries and rational design by evaluating protein sequence datasets and analyzing protein conformational variations [56]. Heinzelman and coworkers showed that SCHEMA structure-guided recombination of three fungal class II cellobiohydrolases (CBHII) has generated highly thermostable CBHII chimeras, which had good performance at a temperature of up to 15°C higher than that of the wildtype, achieving quantitative predictions of thermostability [99].

A cocktail of non-complex cellulases presents synergistic effects on crystalline cellulose digestion. In particular, a mechanistic model-based framework for rational design of optimal cellulase mixtures has been developed to describe the activity of a tertiary cellulase mixture comprising CBHI, CBHII, and EGII, in hydrolyzing lignocellulose with various substrate properties [100]. Furthermore, reconstruction of cellulosomal components is a new direction for improving the efficiency of cellulose hydrolysis. Chimeric constructs of cellulosomal components have been initiated and comprehensively studied by Lamed and Bayer, and coworkers. For example, they converted β-glucosidases (BglA) from a free enzyme to a chimeric CohII-fused enzyme (BglA-CohII), and the BglA-CohII was able to retain the activity of the cellobiohydrolase components and to increase the overall degradation of microcrystalline cellulose compared to using a combination of

the cellulosome and the free β-glucosidase [101]. This effect was considered to result from the immediate removal of product inhibition to the cellobiohydrolase components by the fused BglA.

CHAPTER 2

THEORIES AND BACKGROUND

2.1. Molecular dynamics simulations

2.1.1. Force field

Molecular dynamics (MD) simulation uses force fields to define the molecular system under study. The force fields contain mathematical functions and associated parameters that are used to describe the energy of the molecules as a function of their atomic coordinates. Several force fields have been developed to describe proteins, including AMBER [102], CHARMM [103], GROMACS [104], and OPLS-AA [105]. Comparisons of these protein force fields have been made by Guvench and MacKerell [106]. In this work, the CHARMM22 force field with CMAP [107] was used to describe the protein. The CMAP term is an additional optimization of the backbone Φ, ψ dihedral parameters with a grid-based energy. It was added to the original CHARMM22, since the original CHARMM22 was inaccurate in calculating protein backbone conformational energies [107]. The CHARMM36 all-atom carbohydrate force field [108] was used for β-D-glucopyranose and cellooligomers. The general CHARMM27 force field was used to describe imidazole [109]. The TIP3P model served as the water model [110].

2.1.2. Molecular dynamics simulation

The MD simulations in these studies were carried out using CHARMM [111, 112] and AMBER [113]. These two molecular simulation packages are extensively used in simulating biological systems. AMBER tends to give faster performance than CHARMM for large biological molecular systems. The CHAMBER [114] program was used to convert the CHARMM structure and coordinate files into AMBER format.

CHARMM and AMBER both use classical semi-empirical force fields for MD simulations, and they have a high similarity in the form of the potential energy functions, as shown in Eqs. (2.1) and

(2.2). Compared to V_{AMBER}, there are three extra terms in V_{CHARMM}, which are the Urey-Bradley term that calculates the 1-3 bond energy, the four body quadratic improper dihedral term, and the CMAP term. Appropriate modifications have been made within the code of the AMBER MD engines, SANDER and PMEMD, and their parallel versions [114], to enable the calculation of the energy and derivatives corresponding to these additional terms.

Periodic boundary conditions were applied in the MD simulations to simulate the systems realistically. The long-range van de Waals interaction beyond a cutoff distance is negligible and can be ignored, and a smoothing function is typically applied to ensure continuity in the forces [115]. The pairwise electrostatic forces are not negligible beyond the cutoff, and the long-range electrostatic interaction energy was calculated using the particle-mesh Ewald (PME) method [116, 117]. The SHAKE [118] algorithm was used to fix the covalent bonds that involve hydrogen.

$$V_{AMBER} = \sum_{bonds} k_r\left(r - r_{eq}\right)^2 + \sum_{angles} k_\theta\left(\theta - \theta_{eq}\right)^2 + \sum_{dihedrals} \frac{V_n}{2}\left(1 + \cos(n\phi - \gamma)\right)^2 + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}\right] + \sum_{i<j}\left[\frac{q_i q_j}{\epsilon R_{ij}}\right] \quad (2.1)$$

$$V_{CHARMM} = \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} k_\phi(1 + \cos\left(n\phi - \delta\right))^2 + \sum_{Urey-Bradley} k_u(u - u_0)^2 + \sum_{impropers} k_\omega(\omega - \omega_0)^2 + \sum_{\phi,\psi} V_{CMAP} + \sum_{nonbonded} \epsilon\left[\left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{12} - \left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\epsilon r_{ij}} \quad (2.2)$$

2.1.3. Constant temperature and/or constant pressure simulations

The classical MD simulations were initially designed to run in microcanonical (NVE) ensemble. However, the microcanonical ensemble does not represent the conditions under which

most of the real experiments are carried out. In real experiments, the temperature is controlled instead of the energy, corresponding to the canonical ensemble (NVT). It is not possible to directly switch MD simulations from the microcanonical ensemble to the canonical ensemble through mathematical formulations. Alternatively, several methods have been developed to achieve constant temperature under the microcanonical ensemble [119]. Thermostat strategies mainly include the Berendsen weak coupling method [120], the Andersen thermostat [121], the Nosé–Hoover thermostat [122-124], or the Langevin thermostat [125-127]. These thermostats fix the average temperature of the system over the simulation, and concurrently allow for certain temperature fluctuations that satisfy a canonical distribution.

Additionally, as many experiments are performed under constant temperature and pressure conditions, the isothermal-isobaric ensemble (NPT) has been developed for MD simulations. Similar to temperature control that scales velocity, pressure control is achieved by scaling the dimension of the unit cell under periodic boundary condition. The methods used for thermostats have been adapted to develop barostats: the Berendsen weak coupling barostat was developed based on the Berendsen weak coupling thermostat [120]; the Andersen barostat [121], Nosé–Hoover barostat [122, 128, 129], and Parrinello-Rahman barostat [130] were developed based on the Nosé–Hoover thermostat. These two NPT sampling methods are referred to as the Berendsen week coupling method and the extended system method.

The Nosé–Hoover thermostat is the most realistic temperature coupling method, although it is a slow algorithm. The Berendsen weak coupling thermostat is extremely efficient in relaxing a system to the target temperature, though it does not sample any standard statistical mechanical ensemble, giving rise to inaccurate relationships between the fluctuations of observables (i.e. energy and volume) and second derivatives of the thermodynamic potential (i.e. compressibility) [120]. Nevertheless, a comprehensive study of protein simulations demonstrated that observables such as enthalpy and volume and the calculated compressibility obtained from the extended system

method and the Berendsen weak coupling method were within statistical error of each other [131]. In addition, the Langevin thermostat is commonly used in protein simulations, as it gives fast performance when sampling the large systems of protein in explicit water.

CHARMM is capable of implementing the Berendsen weak coupling method and the extended system method, the latter of which uses a Nosé–Hoover thermostat and an Andersen barostat. Additionally, for the extended system method, the equations of motion for the piston can be replaced by a Langevin equation [132]. AMBER allows for a Berendsen thermostat, an Andersen thermostat, and a Langevin thermostat for temperature control, and employs a Berendsen weak coupling barostat for pressure control. The methods of temperature and pressure regulation that has been used in these studies are discussed in detail below.

2.1.3.1. Nosé–Hoover thermostat

The Nosé–Hoover thermostat [122-124] is based on an extended system that is composed of the real system and an added heat bath reservoir. The energy sampling of the extended system $(\tilde{r}, \tilde{P}, \tilde{t})$ satisfies a microcanonical ensemble, but the energy of the real system is not constant, due to the energy flow between the real system and the heat bath. The heat bath is described by a velocity $\dot{\tilde{s}}$, a friction mass Q, and an artificial dynamical variable $\tilde{s}$. Here $\tilde{s}$ is a time-scaling parameter, and the timescale in the extended system is adjusted to be

$$d\tilde{t} = \tilde{s}(\tilde{t})dt \tag{2.3}$$

The coordinates and velocities are recalculated to be:

$$\tilde{\boldsymbol{r}} = \boldsymbol{r}, \qquad \dot{\tilde{\boldsymbol{r}}} = \frac{\dot{\boldsymbol{r}}}{\tilde{s}}, \qquad \tilde{s} = s, \qquad \dot{\tilde{s}} = \frac{\dot{s}}{\tilde{s}} \tag{2.4}$$

The Lagrangian for the extended system is:

$$\mathcal{L} = \sum \frac{m_i}{2} \tilde{s}^2 \dot{\tilde{\boldsymbol{r}}}_i^{\,2} - U(\tilde{\boldsymbol{r}}) + \frac{1}{2} Q \dot{\tilde{s}}^2 - g k_B T_0 ln\tilde{s} \tag{2.5}$$

in which the first and second terms are the kinetic energy minus the potential energy of the real system, and the third and fourth terms are the kinetic energy minus the potential energy of $\tilde{s}$. For the real-time sampling, $g = N_{df}$, whereas for virtual-time sampling, $g = N_{df} + 1$. Here T_0 is the desired temperature. The Lagrangian equations of motion are derived from Eq. (2.5) to be

$$\ddot{\tilde{\boldsymbol{r}}}_i = \frac{\tilde{\boldsymbol{F}}_i}{m_i \tilde{s}^2} - \frac{2\dot{\tilde{s}}\dot{\tilde{\boldsymbol{r}}}_i}{\tilde{s}} \tag{2.6}$$

$$\ddot{s} = \frac{1}{Q\tilde{s}}\left(\sum m_i \tilde{s}^2 \dot{\tilde{\boldsymbol{r}}}_i{}^2 - g k_B T_0\right) \tag{2.7}$$

Alternatively, the Nosé formulation can be converted into the Nosé–Hoover formulation by the introduction of a real-system variable γ,

$$\gamma = \frac{\dot{s}}{s} = \frac{s\, p_s}{Q} \tag{2.8}$$

in which this case, the Lagrangian equations of motion can be written as

$$\ddot{\boldsymbol{r}}_i = \frac{\boldsymbol{F}_i}{m_i} - \gamma\, \mathbf{r}_i \tag{2.9}$$

$$\dot{\gamma} = \frac{-k_B N_{df}}{Q} T(t)\left(\frac{g}{N_{df}}\frac{T_0}{T(t)} - 1\right) \tag{2.10}$$

The Eqs. (2.9) and (2.10) can be discretized and integrated simultaneously over the MD simulation. The variable γ in the Nosé–Hoover formulation is similar to $\dot{\tilde{s}}$ in the Nosé formulation. When γ (or $\dot{\tilde{s}}$) is positive, heat transfers from the real system to the heat bath, and vice versa.

The Nosé–Hoover thermostat provides smooth, deterministic and time-reversible equations of motion, and the temperature fluctuation exhibits an oscillatory pattern. The value of Q should be set carefully, as too large a Q value will lead to slow convergence to a canonical distribution and too small a Q value will cause too frequent temperature oscillation.

2.1.3.2. Andersen barostat

The Andersen barostat is an extended system method [121]. It mimics the action of a piston on a real system by coupling the system to an external variable V, which is the volume of the unit cell. The piston has a "mass" Q and its kinetic energy and potential energy are described as:

$$\mathrm{E}_{V,kin} = \frac{1}{2} Q \dot{V}^2, \quad \mathrm{E}_{V,pot} = P_{md} V \tag{2.11}$$

where V is treated as the coordinate of the piston and P_{md} is the external pressure that acts on the piston. Therefore, the Lagrangian of the system is:

$$\mathcal{L} = \left\{ \sum_i \left(\mathrm{E}_{kin}(\mathbf{r}_i) - \mathrm{E}_{pot}(\mathbf{r}_i) \right) \right\} + \mathrm{E}_{V,kin} - \mathrm{E}_{V,pot} \tag{2.12}$$

The atomic coordinates of the system $\mathbf{r}_i$ can be rescaled to be $\boldsymbol{\rho}_i$. The momentum conjugate to $\mathbf{r}_i$ is $\mathbf{p_i}$, and the momentum conjugate to $\boldsymbol{\rho}_i$ is $\boldsymbol{\mathfrak{p}}_i$. Their relationships are described as

$$\mathbf{r}_i = V^{1/3} \boldsymbol{\rho}_i \tag{2.13}$$

$$\mathbf{p_i} = \boldsymbol{\mathfrak{p}}_i / V^{1/3} \tag{2.14}$$

The Lagrangian for the rescaled system becomes:

$$\mathcal{L}_2(\boldsymbol{\rho}^N, \dot{\boldsymbol{\rho}}^N, \mathrm{V}, \dot{\mathrm{V}}) = \frac{1}{2} m V^{2/3} \sum_{i=1}^{N} \dot{\boldsymbol{\rho}}_i \cdot \dot{\boldsymbol{\rho}}_i - \sum_{i<j}^{N} \mu\left(V^{1/3} \boldsymbol{\rho}_{ij}\right) + \frac{1}{2} Q \dot{V} - P_{md} V \tag{2.15}$$

Eq. (2.15) is used to derive the equation of motion for the rescaled system. Finally, the equation of motion for the original system is obtained as Eqs. (2.16), (2.17), and (2.18):

$$\dot{\mathbf{r}}_\mathbf{i} = \frac{\mathbf{P_i}}{m} + \frac{1}{3} \mathbf{r_i} \frac{d\ lnV}{dt} \tag{2.16}$$

$$\dot{\mathbf{P}}_\mathbf{i} = -\sum_{j(\neq i)=1}^{N} \hat{\mathrm{r}}_{\mathrm{ij}}\, \mu'(\mathbf{r_{ij}}) - \frac{1}{3} \mathbf{P_i} \frac{d\ lnV}{dt} \tag{2.17}$$

$$Q\ddot{V} = -P_{md} + \left(\frac{2}{3} \sum_{i=1}^{N} \frac{\mathbf{P_i} \cdot \mathbf{P_i}}{2m} - \frac{1}{3} \sum_{i<j=1}^{N} \mathrm{r}_{\mathrm{ij}}\, \mu'(\mathbf{r_{ij}}) \right) / V \tag{2.18}$$

Where μ' is the derivative of μ and $\hat{r}_{ij}$ is the unit vector in the direction $\mathbf{r_i} - \mathbf{r_j}$.

2.1.3.3. Langevin thermostat

The Langevin thermostat is a stochastic dynamics algorithm that is based on the integration of the Langevin equations of motion (Eq. (2.19)) [125-127]. A stochastic Langevin thermostat applies a damping coefficient γ_i and a stochastic force $\mathbf{R}_i(t)$ to the momenta. γ_i does not depend on position and velocity. The stochastic forces $\mathbf{R}_i(t)$ should satisfy Eqs. (2.20) and (2.21). Eq. (2.20) shows that the force component $\mathbf{R}_{iu}(t)$ along the Cartesian axis u is not correlated with the force component $\mathbf{R}_{jv}(t)$ along the axis v unless $i = j, u = v$, and $t = t'$ are all satisfied, when the mean-square components are $2m_i\gamma_i k_B T_0$. The time average of $\mathbf{R}_i(t)$ is zero.

$$\ddot{\boldsymbol{r}}_i(t) = \frac{\boldsymbol{F}_i(t)}{m_i} - \gamma_i \dot{\mathbf{r}}_i(t) - \frac{\mathbf{R}_i(t)}{m_i} \tag{2.19}$$

$$\langle \mathbf{R}_{iu}(t)\, \mathbf{R}_{jv}(t) \rangle = 2m_i\gamma_i k_B T_0 \delta_{ij}\delta_{uv}\delta(t' - t) \tag{2.20}$$

$$\langle \mathbf{R}_i(t) \rangle = 0 \tag{2.21}$$

Langevin dynamics is non-deterministic and time-irreversible. It has been noticed that the Langevin-thermostatted trajectories initiated in the same potential basin, when using the same random seed number, move toward each other as they evolve and finally reach synchronization in phase space, in spite of having different initial atomistic coordinates and momenta [133, 134]. It is suggested that the random seed number in MD simulations is randomly assigned to avoid this "synchronization" artifact.

2.1.3.4. Berendsen weak coupling barostat

The Berendsen weak coupling barostat [120] is similar to the Berendsen thermostat, with the equation of motion described as:

$$\dot{\mathbf{P}}_\mathbf{i} = \frac{P_{md} - P(t)}{\tau_P} \tag{2.22}$$

where $P(t)$ is the instantaneous pressure, P_{md} is the desired constant pressure, and τ_P is the barostat relaxation time constant. This expression leads to cell dimension variations, so that at each step the cell volume is scaled by a factor η, and the coordinates and cell dimensions of an isotropic system are scaled by $\eta^{1/3}$:

$$\eta(t) = 1 - \frac{\Delta t}{\tau_P}\gamma(P_{md} - P(t)) \tag{2.23}$$

where γ is the isothermal compressibility of the system.

2.2. Free energy calculations

Free energy calculation using MD simulations is very useful to study the preferred direction of various reactions. Several MD-based methods have been developed to calculate binding free energies, including traditional free energy perturbation [135], its extension using umbrella sampling combined with the WHAM method [136], thermodynamic integration [135], a nonequilibrium "fast-growth" method known as Jarzynski's equality [137, 138], and the free energy estimation for host-guest type of binding interaction using the calculated volume density map [139].

The studies reported in Chapter 3 employed the theory of hydrophobicity [140] to calculate the free energy cost of small cavity formation in protein structures. The studies in Chapter 5 adopted steered molecular dynamics simulations combined with Jarzynski's equality to calculate cellobiose product expulsion energies in family 48 cellulases and their rational mutants. The studies described in Chapter 6 used the calculated volume density map to estimate the binding free energy between the planar molecules imidazole-imidazole, and imidazole-glucose in aqueous solution.

2.3. Background of Family 48 Cellulases

This thesis is focused on studying the structure-function relationship of family 48 cellulases

and on improving the cellulases from theoretical perspectives. Family 48 cellulases (GH48) are a major group of processive exocellulases that catalyze cellulose degradation into primarily cellobiose. More than twenty GH48s have been identified from various microorganisms [32]. The X-ray crystal structures of five GH48s have been solved, which are Cel48 from *Bacillus pumilus* (Markus Alahuhta, personal communication), CelA from *Caldicellulosiruptor bescii* [141], CelS from *Clostridium thermocellum* [142], CelF from *Clostridium cellulolyticum* [143-145], and Cel48A from *Thermobifida fusca* (Markus Alahuhta, personal communication). The crystal structures show that GH48s are globular proteins, and that they share common features including an $(\alpha/\alpha)_6$ barrel structure that points to the active site from protein surface and an active site tunnel (Figure 2.1). The active site tunnel contains multiple Trp residues, which assists in its processive action on the cellulose. Besides, it features an aromatic zone composed of five aromatic residues around the substrate at subsites -2 and -3. This structural feature might be used to strip the water molecules around the cellulose chain and position it correctly to prepare for the hydrolytic event. The active site tunnel provides seven substrate subsites preceding the hydrolytic cleavage site and two after it at the tunnel exit, serving as a substrate pathway for processive action. These subsites are named as subsites -7, -6, -5, -4, -3, -2, -1, +1, and +2 from the substrate's nonreducing end at the tunnel entrance to the reducing end at the tunnel exit (Figure 2.2). It is generally believed that family 48 cellulases can recognize the cellulose chains by their reducing end, and acquire them into the active site tunnel. Subsequently, the cellulose chain progresses through the tunnel until it is in position for hydrolytic reaction. Family 48 cellulases follow an inverting mechanism [28, 32, 146]. In particular, they take advantage of a catalytic acid (glutamic acid) and a catalytic base (aspartic acid) to achieve the hydrolysis of glycosidic bonds in cellulose chains. As a result, mostly a cellobiose product is cleaved off and released to the aqueous environment. Next, the cellulose chain progresses forward in the tunnel by a cellobiose unit so that the catalytic cycle is continued. The cellulases can dissociate from the cellulose substrate, halting the processive hydrolysis.

The enzymatic activities of family 48 cellulases are reported to be extremely low. *T. fusca* Cel48A exhibited an activity of 0.07~0.4 μmol cellobiose · $minute^{-1}$ · μmole $enzyme^{-1}$ for various cellulosic substrates [147]. The melting temperatures for Cel48, CelF, CelS, and CelA are 45 ^{o}C, 55 ^{o}C, 65 ^{o}C, and 85 ^{o}C, respectively (Yannick J. Bomble, personal communication). With respect to enzymatic activity, Cel48, CelF, and Cel48A favor mesophilic conditions, CelS is thermophilic, and CelA is extremely thermophilic since it exhibits optimal activity at 75 ^{o}C and it sustains activity at high temperature up to 90 ^{o}C [148]. At optimal conditions, CelA has the relatively highest activity, and Cel48 has the lowest activity (Yannick J. Bomble, personal communication). The thermophilic cellulases are particularly interesting in that they can be added directly to cellulosic biomass immediately after pretreatment under high temperature conditions, increasing energy efficiency. In addition, CelS and CelF are critical components of cellulosomes, which seem to be better alternatives to simple combinations of synergistic cellulases for biomass degradation [149].

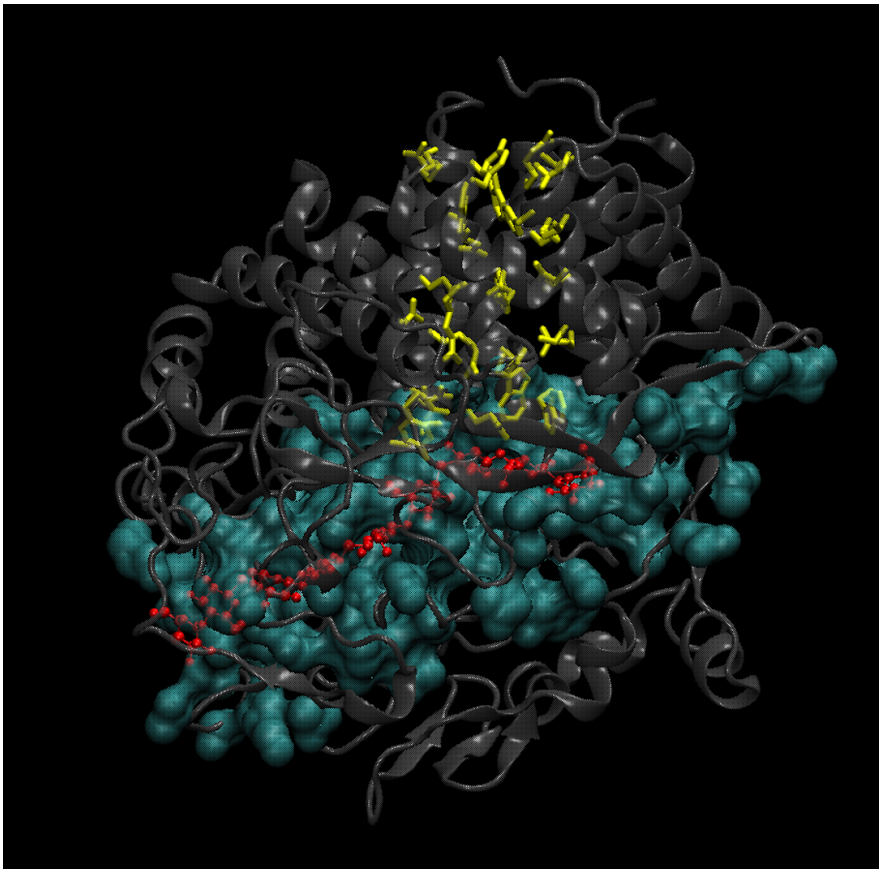


Figure 2.1. A representative of GH48s using the crystal structure of CelF_E55Q (2QNO.pdb). Certain amino acid residues on the six inner α-helices of the $(\alpha/\alpha)_6$ barrel structure (yellow sticks) form a pore structure that connects the protein surface to the active site. An active site tunnel (rendered in cyan) accommodates a cellooligomer substrate ligand (red ball-and-sticks).

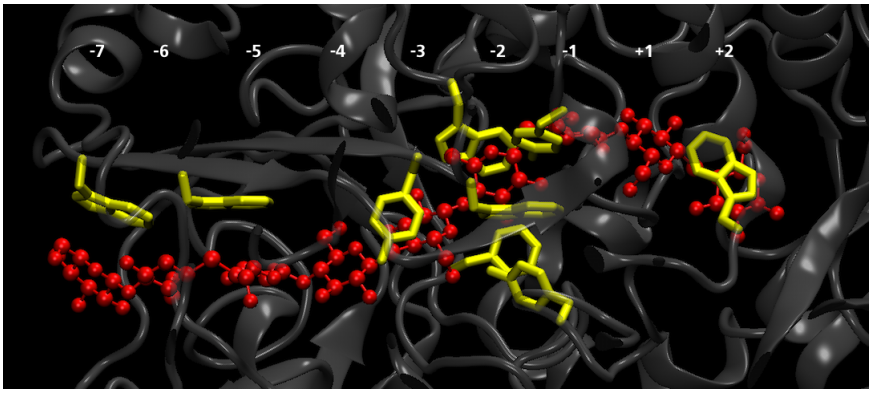


Figure 2.2. The active site tunnel of GH48s is composed of multiple Trp's along the tunnel and an aromatic zone near subsite -2 and -3 (yellow sticks). The cellooligomer substrate is shown in red ball-and-sticks.

CHAPTER 3

MOLECULAR MODELING AND SIMULATIONS OF A UNIQUE WATER-FILLED PORE STRUCTURE WITH POSSIBLE MECHANISTIC IMPLICATIONS IN FAMILY 48 CELLULASES

3.1. Introduction

Lignocellulosic biomass is commonly considered as a promising resource for producing renewable bioenergy in the form of ethanol, though it is extremely difficult to deconstruct. The degradation of lignocellulose involves biomass pretreatment, cellulose decomposition by hydrolytic enzymes, and sugar fermentation. In this process, the cellulose solubilization catalyzed by cellulases is the rate-limiting step. Understanding the mechanism of cellulose hydrolysis by cellulases is crucially important, and might further benefit the improvement of biomass conversion process for biofuel production.

Family 48 glycoside hydrolases are a group of the most important processive exocellulases. They catalyze the hydrolysis of cellulose chains from the reducing ends, producing predominantly cellobiose, a small amount of cellotriose, and a trace amount of cellotetraose. They exist either as the catalytic domains of noncomplex cellulases, such as Cel48A from *Thermobifida fusca*, or as the catalytic components in cellulosomes, such as CelS from *Clostridium thermocellum* and CelF from *Clostridium cellulolyticum*. As cellulosomal components, family 48 cellulases are of key importance in crystalline cellulose hydrolysis [150, 151], and they are considered promising candidates for "designer cellulosomes" [152].

Family 48 cellulases function via an inverting mechanism [28, 32, 146]. They acquire a cellulose chain from its reducing end, feed it into an active site tunnel in the globular protein until it reaches the active site asymmetrically locating at the product side of the tunnel. The active site of family 48 cellulases employs a pair of amino acids, namely a glutamic acid as the catalytic acid and an aspartic acid as the catalytic base, to achieve the hydrolysis. Prior to substrate hydrolysis, the

aspartic acid sidechain carboxyl group is unprotonated, forming a nucleophilic complex with a water molecule. This nucleophilic complex attacks the anomeric carbon of the substrate unit at the -1 subsite to trigger the hydrolysis. On the other side, the glutamic acid carboxyl sidechain is protonated, and attracts the glycosidic oxygen to gradually form a covalent bond, ending up breaking the glycosidic bond between two 1,4-β-D-glucosyl units and resulting in an inversion of the stereochemistry at the product cleavage site. Water is critical to the hydrolysis since each catalytic cycle consumes a water molecule (Figure 3.1).

The structures of several family 48 cellulases have been solved by X-ray crystallography, and they present unique structural features of the family 48 cellulases. In particular, they contain a very unique water-filled pore structure that is formed by the six inner α-helices of a characteristic $(\alpha/\alpha)_6$ barrel structure. This water pore structure is approximately 26 Å in length, connecting the protein surface to the active site at the core of the protein. There are 9 crystallized water molecules present within the water pore region of the *C. thermocellum* CelF crystal structures [144] and 12 water molecules within the pore of the *T. fusca* Cel48A crystal structure (Markus Alahuhta, personal communication), suggesting that this pore might function as a channel for water transport (Figure 3.2). Interestingly, many studies have described water molecules that form "wires", or clusters in nonpolar confinement that are important in protein function [153].

It is possible that the water pore structure formed by the α-helices provides a pathway for water transport to the active site, particularly since that the catalytic residues reside at one end of water pore structure and that there are water molecules lining up within the pore of the crystal structures. Hence, we hypothesized that the water pore structure of family 48 cellulases might be of mechanistic importance in transporting water molecules into the active site for substrate hydrolysis.

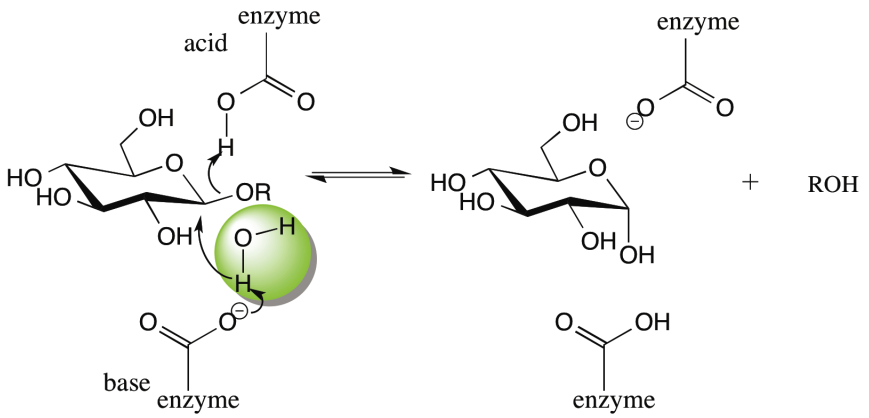


Figure 3.1. Inverting mechanism of family 48 cellulases.

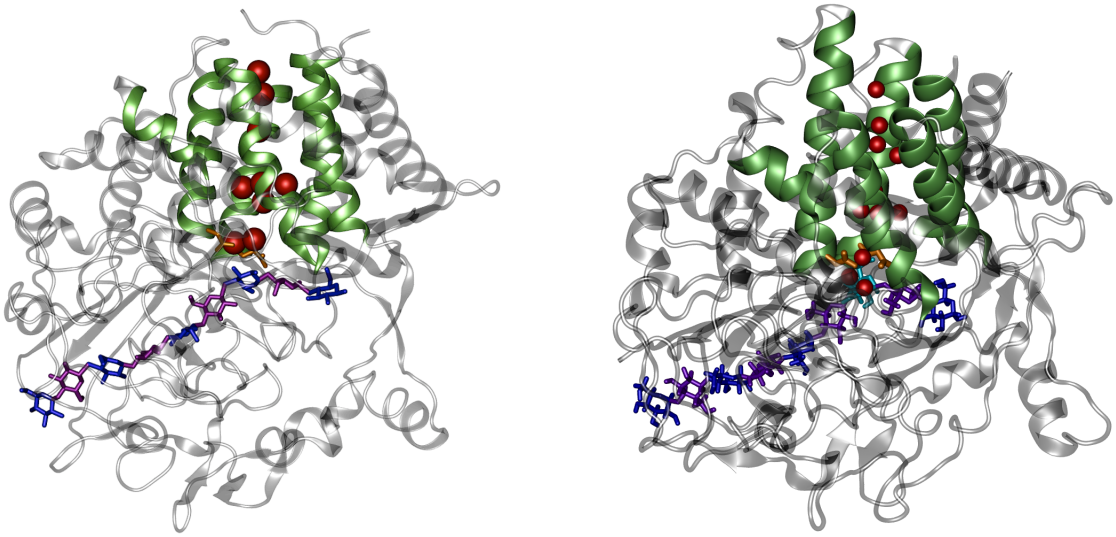


Figure 3.2. Cartoon representation of the crystal structures of CelF (left) and Cel48A (right). The ligand is illustrated with successive glucose residues alternately colored dark blue and purple. The ligand in CelF is a pseudo-substrate, hemithiocellooligosaccharides. The ligand in Cel48A is composed of a heptamer and a dimer with the -1 subsite missing, and it has been built with energy minimization using CHARMM and in this illustration, it is displayed in light blue. The catalytic acid and base are shown in orange. The helices constituting the hypothesized water pore structure are shown in green, and the crystallographic water molecules that fill this pore are shown as red van der Waals spheres.

3.2. Methods

The hypothesis could be proved if we are able to construct a mutant of family 48 cellulases in which the water passage through the water pore structure is blocked, and if the catalytic activity of such a mutant was significantly reduced. To test this hypothesis, we have employed molecular modeling and simulations, which this chapter covers, combined with experimental studies that were conducted by Maxim Kostylev under the guidance of Dr. David Wilson.

3.2.1. Three-dimensional structures of family 48 cellulases

We used the X-ray crystal structures of family 48 cellulases for the modeling and molecular simulations. Several X-ray crystal structures of CelF with various ligands taking up the substrate and/or the product site in the protein have been solved and reported by Parsiegla and coworkers [143-145]. The most recently reported ones correspond to two CelF single mutants, namely CelF_E55Q and CelF_E44Q, with each containing a substrate analog that fills the active site tunnel [7]. The former contains a thio-oligosaccharide nonamer taking up the positions from the -7 to +2 subsites of a "lower path" in the active site tunnel of the enzyme, while the latter contains a thio-oligosaccharide octamer occupying the -7 to +1 subsites of an "upper path" in the tunnel. Compared with the crystal structure of another family 48 cellulase, *C. thermocellum* CelS [142], the CelF_E55Q structure seems to represent the CelF structure more realistically, and therefore was used to construct the wildtype CelF.

In addition, we studied another family 48 cellulase, *T. fusca* Cel48A, which was available to us experimentally from Dr. David Wilson's lab collections. The structure of Cel48A was built by homologous modeling, since its crystal structure was not available until recently. To acquire the optimal template for building the Cel48A homologous model, the Standard Protein BLAST engine with the Position-Specific Iterated BLAST algorithm was employed to perform Cel48A sequence alignment to all the protein crystal structures available at the Protein Data Bank [154]. As a result,

the CelF_E55Q structure was found the most suitable template (Table 3.1), and the homologous model was created by Swiss-PdbViewer using the "magic fit" aligning algorithm [155]. Comparison of the Cel48A crystal structure and its homologous model showed that they have minor differences in their three-dimensional structures and the residues with large RMSD difference between the two structures locate mainly in flexible coils, which fluctuate easily in MD simulations (Figure 3.3).

3.2.2. Molecular dynamics simulation

Molecular dynamics (MD) simulations were used to study the behavior of cellulases and their mutants in aqueous solution. The MD program CHARMM [111, 112] was used to build the molecular systems. The cellulase molecule with a cellooligomer ligand (DP=9) taking up the subsites -7 to +2 was built from the crystal structure or the homologous model. The protein and ligand were placed in a cubic water box with the dimension of ~90Å, and the overlapping water molecules were removed, ending up with about 20000 water molecules in each system. The sodium counter ions were added to the solution to neutralize the charges of the systems. The CHARMM22/CMAP protein force field [103, 107], the all-atom carbohydrate force field [108], and the TIP3P model [110] were used to describe the protein, the cellooligomer, and the water molecules, respectively. The parallel version of PMEMD engine of AMBER [113] was employed to carry out the MD simulations, for which the structure, coordinate, and force filed files in CHARMM format were converted into the AMBER format using CHAMBER [114]. The SHAKE algorithm [118] was applied in the simulation to constrain the bond distance involving hydrogen atoms. The nonbond cutoff distance was 8Å.

The system preparation of Cel48A and its mutants followed four stages, which were: solvent minimization with 1000 steepest descent steps and 1000 conjugate gradient steps; system minimization with the same minimization strategy; equilibration of the solvent at a constant volume from 0 K to 300 K for 20 ps; and equilibration of the system at a constant temperature of

300 K using a Langevin thermostat and at constant pressure of 1 atm using a Berendsen weak coupling barostat for 500 ps with a step size of 2 fs. The production run was carried out under constant temperature and pressure conditions with the same strategy as that for the system equilibration, and was initially carried out for 24 ns. In particular, for the Cel48A wild type and Mutant C the production runs were carried out three times, with the water box positioned at different orientations for each run, and each production run was carried out for 100 ns.

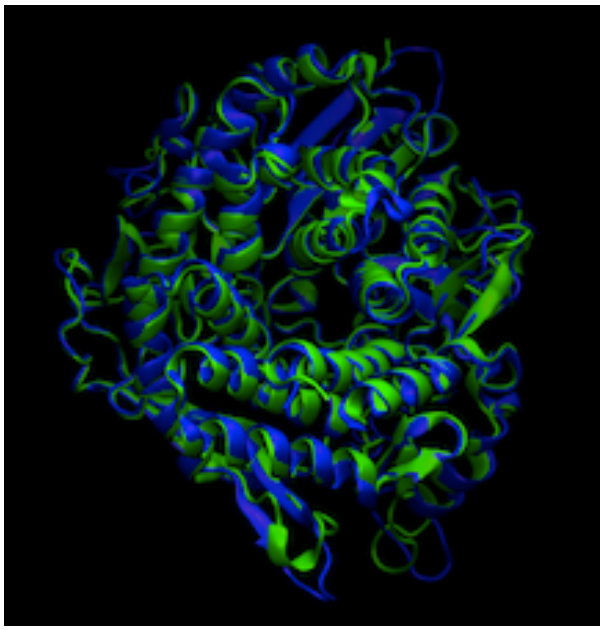
For the CelF water pore mutant, the "dewetting" and "wetting" simulations were carried out using the same system preparation strategy as that for Cel48A except that the equilibration of the system was conducted for 100 ps. Next, the production run of the CelF mutant were conducted for 400 ps.

3.2.3. Identification of water pore residues and selection of pore waters

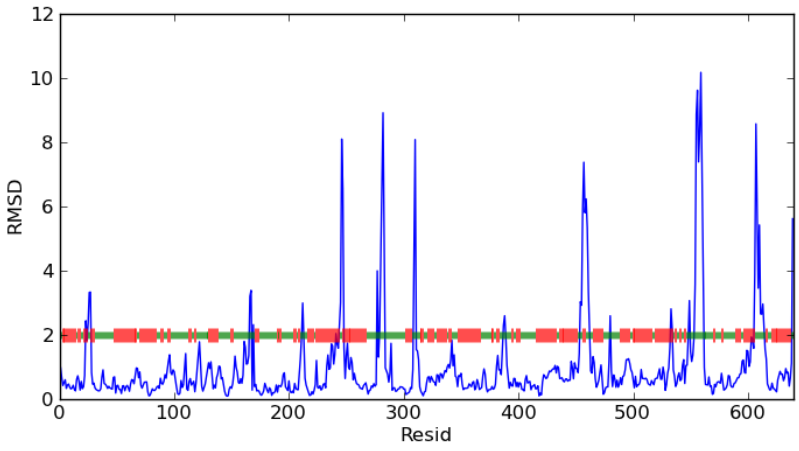
The amino acid residues composing the water pore structures of three family 48 cellulases, CelF, CelS, and Cel48A, respectively, were identified using VMD [156]. The alignment of the water pore amino acid sequences showed that 15 out of the 31 water pore residues were conserved and 12 were partially conserved, suggesting significant amino acid sequence conservation for this structure (Table 3.2). Conservation of protein sequence is often the result of structure-function relationships, providing a foundation for our hypothesis.

To analyze water flow within the water pore structure, the structure was divided into five sections, termed Rings 1 to 5 (Figure 3.4). These ring-shaped sections stacked together, forming a column-like shape connecting the active site to the protein surface. The components of each Ring in CelF and Cel48A are presented in Table 3.3. To select the water molecules within each ring for all frames of the trajectories, the trajectory was loaded in VMD and VMD Tcl scripting was used to orient the system of the first frame such that the primary axis of the selected ring residues aligned to the X, Y, and Z directions. Then, the whole trajectory was aligned to the selected ring residues in

the first frame. The criteria of water selection within each ring were defined as shown in Table 3.3, and the water selection was performed accordingly for the whole trajectory to obtain the residue IDs of the selected water molecules over time, which allowed for quantitative analysis of water flow in the pore.



(a)



(b)

Figure 3.3. (a) Superimposition of Cel48A crystal structure (in green) and the Cel48A homologous model (in blue); (b) Residue-by-residue RMSD difference between Cel48A crystal structure and the homologous model (in blue). The red and green markers show the secondary structure of the protein, with the red corresponding to the more rigid secondary structures (such as α-helices, β-sheets, 3-10 helices, and Pi helices) and the green corresponding to flexible coils.

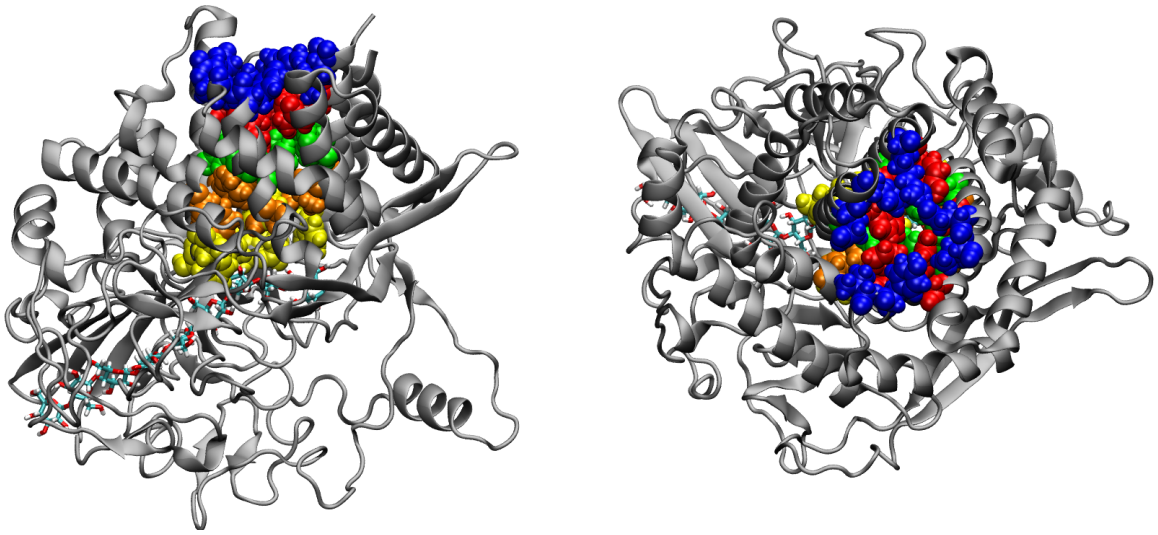


Figure 3.4. Side view (left) and top view (right) of the water pore structure in Cel48A. Ring 1, 2, 3, 4, and 5 were colored in blue, red, green, orange, and yellow.

Table 3.1. Protein sequence alignment of Cel48A.

Chain A, Crystal Structure Of The Mutant E55q Of The Cellulase Cel48f In Complex With A Thio-Oligosaccharide
Sequence ID: pdb|2QNO|A

Alignment statistics for match #1

Score	Expect	Method	Identities	Positives	Gaps
657 bits(1694)	0.0	Compositional matrix adjust.	350/643(54%)	428/643(66%)	31/643(4%)

```
Query  2    YDQAFLEQYEKIKDPASGYFREFNGLLVPYHSVETMIVEAPDHGHQTTSEAFSYYLWLEA  61
            Y   F   Y KIKDPA+GYF E     +PYHS+ET++VEAPD+GH TTS+A SYY+WLEA
Sbjct  9    YQDRFESMYSKIKDPANGYFSEQG---IPYHSIETLMVEAPDYGHVTTSQAMSYYMWLEA  65

Query  62   YYGRVTGDWKPLHDAWESMETFIIPGTKDQPTNSA--YNPNSPATYIPEQPNADGYPSPL  119
             +GR +GD+     +W   E ++IP  KDQP  S   Y+ N PATY PE  +   YPSPL
Sbjct  66   MHGRFSGDFTGFDKSWSVTEQYLIPTEKDQPNTSMSRYDANKPATYAPEFQDPSKYPSPL  125

Query  120  MNNVPVGQDPLAQELSSTYGTNEIYGMHWLLDVDNVYGFGFCGDGTDDAPAYINTYQRGA  179
              + PVG+DP+  +L+S YGT+ +YGMHW+LDVDN YGFG   DGT   P+YINT+QRG
Sbjct  126  DTSQPVGRDPINSQLTSAYGTSMLYGMHWILDVDNWYGFGARADGTSK-PSYINTFQRGE  184

Query  180  RESVWETIPHPSCDDFTHGGPNGYLDLFTDDQNY-AKQWRYTNAPDADARAVQVMFWAHE  238
            +ES WETIP P  D+   GG  G+LDLFT D    AKQ++YTNAPDADARAVQ  +WA +
Sbjct  185  QESTWETIPQPCWDEHKFGGQYGFLDLFTKDTGTPAKQFKYTNAPDADARAVQATYWADQ  244

Query  239  WAKEQGKENEIAGLMDKASKMGDYLRYAMFDKYFKKIGNCVGATSCPGGQGKDSAHYLLS  298
            WAKEQGK   ++  + KA+KMGDYLRY+ FDKYF+KIG    A     G G D+AHYLLS
Sbjct  245  WAKEQGKS--VSTSVGKATKMGDYLRYSFFDKYFRKIGQPSQA-----GTGYDAAHYLLS  297

Query  299  WYYSWGGSLDTSSAWAWRIGSSSSHQGYQNVLAAYALSQVPELQPDSPTGVQDWATSFDR  358
            WYY+WGG +D  S W+W IGSS +H GYQN  AA+ LS     +P S  G  DWA S DR
Sbjct  298  WYYAWGGGID--STWSWIIGSSHNHFGYQNPFAAWVLSTDANFKPKSSNGASDWAKSLDR  355

Query  359  QLEFLQWLQSAEGGIAGGATNSWKGSYDTPPTGLSQFYGMYYDWQPVWNDPPSNNWFGFQ  418
            QLEF QWLQSAEG IAGGATNSW G Y+  P+G S FYGM Y   PV+ DP SN WFG Q
Sbjct  356  QLEFYQWLQSAEGAIAGGATNSWNGRYEAVPSGTSTFYGMGYVENPVYADPGSNTWFGMQ  415

Query  419  VWNMERVAQLYYVTGDARAEAILDKWVPWAIQHTDVDADNGGQNFQVPSDLEWSGQPDTW  478
            VW+M+RVA+LYY TGDARA+ +LDKW  W       +AD     FQ+PS ++W GQPDTW
Sbjct  416  VWSMQRVAELYYKTGDARAKKLLDKWAKWINGEIKFNADG---TFQIPSTIDWEGQPDTW  472

Query  479  TGT--YTGNPNLHVQVVSYSQDVGVTAALAKTLMYYAKRSGDTTALATAEGLLDALL-AH  535
              T  YTGN NLHV+VV+Y  D+G  ++LA TL YYA +SGD T+   A+ LLDA+   +
Sbjct  473  NPTQGYTGNANLHVKVVNYGTDLGCASSLANTLTYYAAKSGDETSRQNAQKLLDAMWNNY  532

Query  536  RDSIGIATPEQPS-WDRLDDPWDGSEGLYVPPGWSGTMPNGDRIEPGATFLSIRSFYKND  594
             DS GI+T EQ   + R  D     + ++VP GW+G MPNGD I+ G  F+ IRS YK D
Sbjct  533  SDSKGISTVEQRGDYHRFLD-----QEVFVPAGWTGKMPNGDVIKSGVKFIDIRSKYKQD  587

Query  595  PLWPQVEAHLNDPQNVPAPIVERHRFWAQVEIATAFAAHDELF  637
            P W  + A L   Q    P    HRFWAQ E A A   +  LF
Sbjct  588  PEWQTMVAALQAGQ---VPTQRLHRFWAQSEFAVANGVYAILF  627
```

Table 3.2. The amino acid sequence of water pore residues in three family 48 cellulases.

Location*	Cel48A		CelF		CelS	
	Name	ID	Name	ID	Name	ID
Helix 1	SER	50	SER	54	SER	86
	GLU	51	GLU	55	GLU	87
	TYR	55	TYR	59	TYR	91
	TRP	58	TRP	62	TRP	94
	TYR	62	MET	66	MET	98
	ARG	65	ARG	69	ASN	101
Helix 2	ALA	222	ALA	228	ALA	253
	ASP	224	ASP	230	ASP	255
	ALA	225	ALA	231	ALA	256
	ARG	228	ARG	234	ARG	259
	GLN	231	GLN	237	GLN	262
	TRP	235	TRP	241	TRP	266
	GLU	238	GLN	244	LYS	269
	TRP	239	TRP	245	TRP	270
Helix 3	TYR	326	TYR	323	TYR	351
	ASN	328	ASN	325	ASN	353
Helix 4	TRP	420	TRP	417	TRP	445
	GLU	423	GLN	420	GLN	448
	ARG	424	ARG	421	ARG	449
	GLN	427	GLU	424	GLU	452
	TYR	430	TYR	427	LEU	455
	VAL	431	LYS	428	GLU	456
Helix 5	ALA	504	SER	500	SER	526
	LYS	507	ASN	503	ASN	529
	TYR	511	TYR	507	THR	533
Helix 6	TRP	621	TRP	611	TRP	645
	GLU	625	GLU	615	ASP	649
	THR	628	VAL	618	VAL	652
	ALA	632	VAL	622	VAL	656
	GLU	635	ILE	625	THR	659
	LEU	636	LEU	626	TYR	660

Note:
Residues in blue are conserved residues among the three family 48 cellulases.
Residues in red are partially conserved residues.
* .The inner six α-helices that form the water pore structure.

Table 3.3. The residues corresponding to each ring of CelF and Cel48A, and pore water selection.

Ring	CelF wildtype	Cel48A wildtype	Water selection criteria (Å)
1	ARG69, GLN244, LYS428, TYR427, ILE625, LEU626, (TRP245 and not atom HE3 HZ3 CZ3 CH2 CZ2 CE3 HH2)	ARG65, GLU238, VAL431, TYR430, GLU635, LEU636, (TRP239 and not atom HE3 HZ3 CZ3 CH2 CZ2 CE3 HH2)	Xmax - 1.0 > X > Xmin - 0.4 Ymax - 1.2 > Y > Ymin + 1.2 Zmax - 1.2 > Z > Zmin + 1.2 (2*Y/(Ymax-Ymin))^2 + (2*Z/(Zmax-Zmin))^2 <=1
2	TYR507, TRP241, VAL622, MET66, GLU424, (TRP245 and atom HE3 HZ3 CZ3 CH2 CZ2 CE3 HH2)	TYR511, TRP235, ALA632, TYR62, GLN427, (TRP239 and atom HE3 HZ3 CZ3 CH2 CZ2 CE3 HH2)	Xmax + 1.0 > X > Xmin – 1.0 Ymax - 1.2 > Y > Ymin + 1.2 Zmax - 1.2 > Z > Zmin + 1.2 (2*Y/(Ymax-Ymin))^2 + (2*Z/(Zmax-Zmin))^2 <=1
3	VAL618, ASN503, TRP62, GLN420, GLN237, (ARG421 and backbone atoms)	THR628, LYS507, TRP58 GLU423, GLN231, (ARG424 and backbone atoms)	Xmax + 1.0 > X > Xmin – 1.0 Ymax - 1.2 > Y > Ymin + 1.2 Zmax - 1.2 > Z > Zmin + 1.2 (2*Y/(Ymax-Ymin))^2 + (2*Z/(Zmax-Zmin))^2 <=1
4	ASN325, SER500, TYR59, ARG234 GLU615, (TRP417 and atom C CA N O HB2 CB HB1 HD1 CD1 NE HE1), (ARG421 and sidechain atoms)	ASN328, ALA504, TYR55, ARG228, GLU625, (TRP420 and atom C CA N O HB2 CB HB1 HD1 CD1 NE HE1), (ARG424 and sidechain atoms)	Xmax + 1.0 > X > Xmin – 0.5 Ymax - 1.2 > Y > Ymin + 1.2 Zmax - 1.2 > Z > Zmin + 1.2 (2*Y/(Ymax-Ymin))^2 + (2*Z/(Zmax-Zmin))^2 <=1
5	TRP611, ASP230, GLU55, ALA231, TYR323, SER54, ALA228, (TRP417 and sidechain atoms except atom HB2 CB HB1 HD1 CD1 NE HE1)	TRP621, ASP224, GLU51 ALA225, TYR326, SER50, ALA222, (TRP420 and sidechain atoms except atom HB2 CB HB1 HD1 CD1 NE HE1)	Xmax + 1.0 > X > Xmin +0.4 Ymax - 1.2 > Y > Ymin + 1.2 Zmax - 1.2 > Z > Zmin + 1.2 (2*Y/(Ymax-Ymin))^2 + (2*Z/(Zmax-Zmin))^2 <=1

Note:
The positive side of X-axis is set to point from the active site to the protein surface, and vise versa.
The Cartesian coordinate (X,Y,Z) describes center of geometry of water molecules.
The residue selections are following the VMD selection routine.

3.3. Results and Discussions

3.3.1. Rational design of cellulase mutants

To find mutants that would prevent the movement of water through the pore, initially an attempt was made to mechanically block the pore by substitution with bulky side chains and with the intention of stabilizing the mutant water pore structure. Two single-site mutants of the CelF enzyme were made: Q420W, replacing Gln420 with a Trp residue in Ring 3 of the pore, and W62R, replacing a wild type Trp in Ring 3 with an Arg, which was anticipated to make salt bridges across the pore with Glu424 and Glu615. Three double substitutions introducing bulky Trp side chains were also attempted: Q420W and V622W, Q420W and V618W, and Q420W and Q237W, where in each case the pair was chosen such that the selected residues were on opposite sides of the pore. Each of these mutants was simulated for no more than 2 ns, since none of these substitutions was found to inhibit the diffusion of water through the pore, as the bulky side chains simply rotated out of the way on a frequent basis. From these results it seemed unlikely that the pore could be closed off solely by mechanical obstruction.

Converting the water pore to be sufficiently hydrophobic served as an alternative approach for pore blocking [157, 158]. We constructed a CelF mutant by substituting all the water pore residues except the catalytic acid (Glu55) and catalytic base (Asp230) with Ile's. With the same strategy, a series of Cel48A mutants were designed with a reduced number of mutation sites to ease the experimental efforts. The mutations sites were all located within the inner Rings (Ring 2, 3, and 4) of the water pore (Table 3.5 and Figure 3.5), and the residues at the mutation sites of Cel48A were converted into Phe, a bulky and hydrophobic amino acid residue.

Table 3.5. The mutation sites of Cel48A water pore mutants.

Location	Mutant 1	Mutant 2	Mutant 3	Mutant 4	Mutant C	Ring Site
Helix 1		GLU423 ARG424	GLN427 GLU423 ARG424	GLN427 GLU423 ARG424	GLN427 * GLU423 * ARG424 #	Ring 2 Ring 3 Ring 3, 4
Helix 2	LYS507	LYS507	LYS507	LYS507	LYS507 *	Ring 3
Helix 3	ASN328	ASN328	ASN328		ASN328 #	Ring 4
Helix 4				GLN231 ARG228	GLN231 # ARG228 #	Ring 3 Ring 4
Helix 5	TYR55	Tyr55	TYR55		TYR55 #	Ring 4
Helix 6	GLU625	Glu625	GLU625	GLU625	GLU625 *	Ring 4

Note: #. Conserved residue in CelF, CelS, and Cel48A.
*. Partially conserved residue in CelF, CelS, and Cel48A.

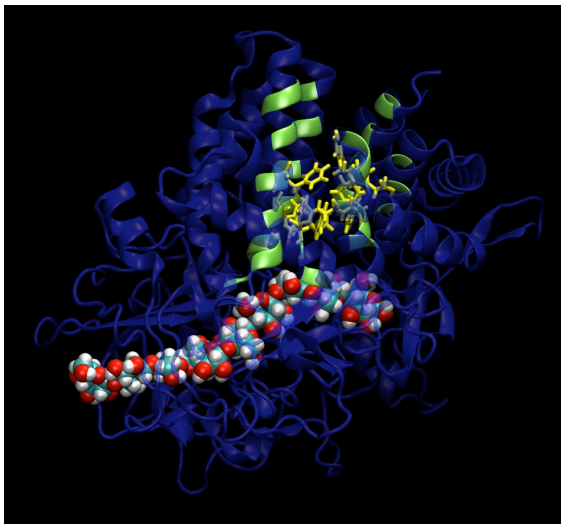
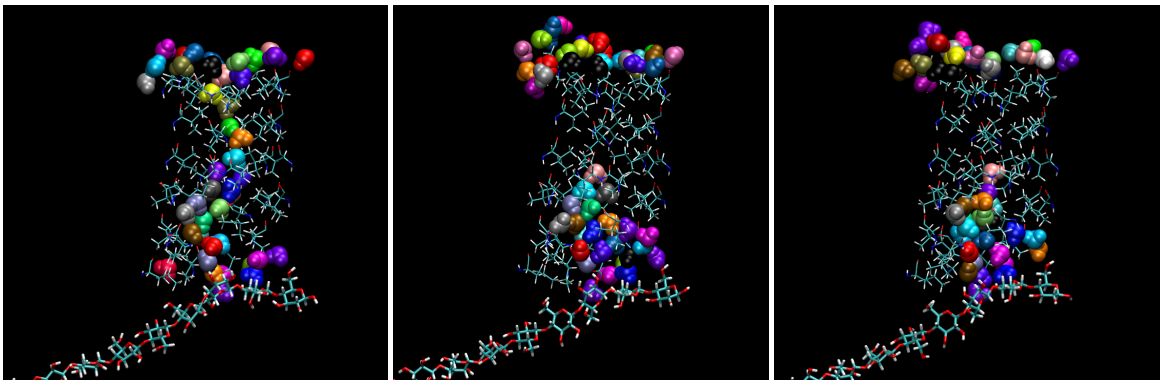


Figure 3.5. Mutation sites of Mutant C, the Cel48A mutant with the largest number of mutation sites.

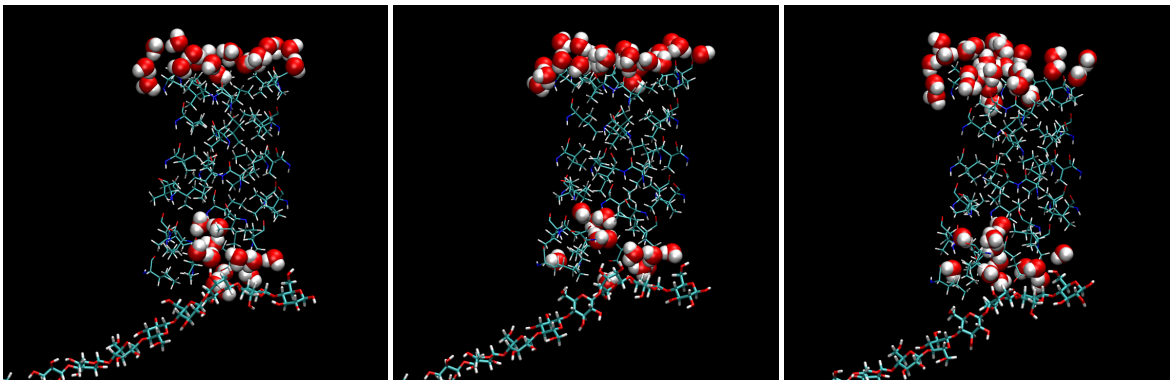
3.3.2. CelF mutant

With regard to the CelF mutant, the "dewetting" simulation was initiated by filling the water pore with water. After about 200 ps, the water molecules were completely depleted (Figure 3.6a). On the other hand, the "wetting" simulation evacuated the water molecules from the water pore at the beginning stage, and no water molecules filled the pore after ~100 ps (Figure 3.6b). The results indicated that hydrophobic effect could be applied to block water passage through the water pore structure.



t=0 t=200 ps t=400 ps

(a)



t=0 t=200 ps t=400 ps

(b)

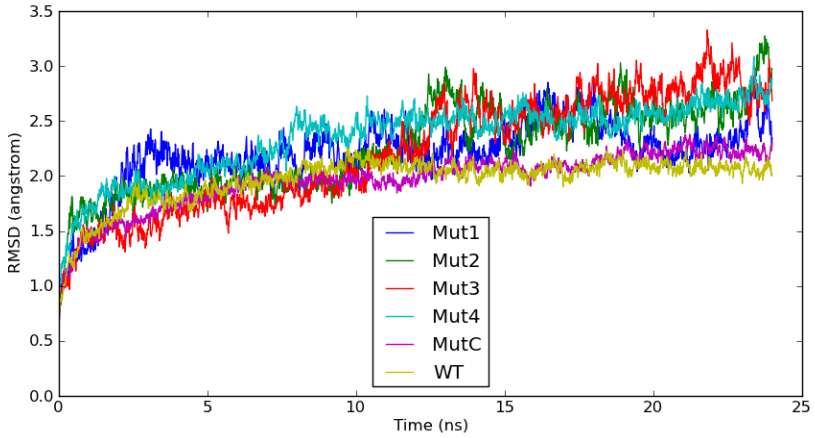
Figure 3.6. "Dewetting" (a) and "wetting" (b) simulations of CelF water pore mutant, in which all water pore residues were converted into Ile's except Glu55 and Asp230. The water molecules within or near the water pore are shown using VdW spheres; the water pore residues are shown using thin licorice; and the substrate residues are shown in medium licorice.

3.3.3. Cel48A mutants

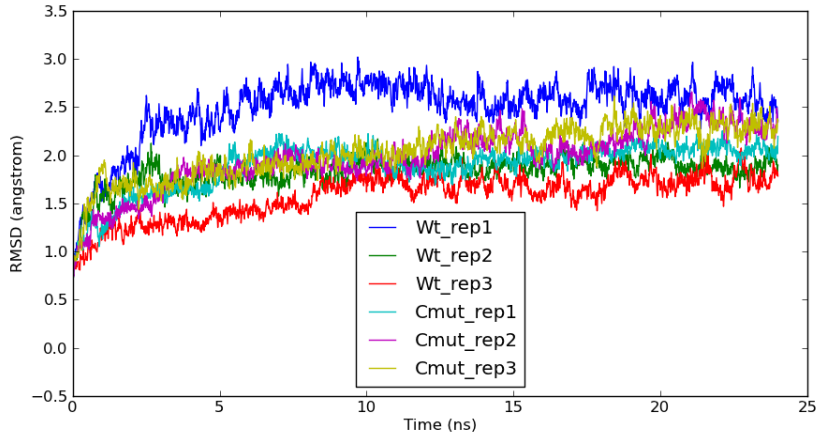
3.3.3.1. RMSD

While short-term (~100 ns) MD simulations cannot determine whether a protein conformation is stable, no tendency for protein unfolding was observed for the Cel48A wildtype and mutants within the simulation time scale. In each case, the trajectory RMSD of the protein backbone atoms (C, CA, and N) was calculated to be within the range of 1.5Å to 3.0Å, setting the briefly equilibrated protein structure as a reference frame (Figure 3.7). Because of insufficient

sampling of the protein structure ensembles, we could not conclude if the mutants would succeed in folding properly, or if the mutants could maintain a stable conformation in the long term. The movement of water in the water pore illustrated that Mutant C was the best candidate among all the limited mutants in preventing water transport through the pore.



(a)



(b)

Figure 3.7. (a) Trajectory RMSDs of Cel48A wildtype and mutants, in which the RMSDs of wildtype (WT) and Mutant C (MutC) were average values of the three production runs. (b) Trajectory RMSDs of Cel48A wildtype (WT) and Mutant C (Cmut). "rep1", "rep2", and "ref3" referred to the three repetitions of the simulations.

3.3.3.2. Water motions in the pore as a function of time

The movement of water molecules in the water pore was quantified. The residue IDs of the water molecules within each ring of the water pore structure were collected over the simulation time. The temporal evolution of the number of water molecules within the inner rings (Ring 2, Ring 3, and Ring 4) demonstrated that almost no water molecules were present in Ring 3 of Mutant C, and the total number of water molecules within the three inner Rings in Mutant C was much less than that in the wildtype (Figure 3.8). Further, all the water molecules that had occurred in the pore were identified and their positions were traced in the water pore over the simulation time, and it was observed that water molecules could move from one side to the other of the water pore in the wildtype, but could not go through the pore in Mutant C. This result demonstrated that the hydrophobic effect was capable of blocking water transport through the water pore.

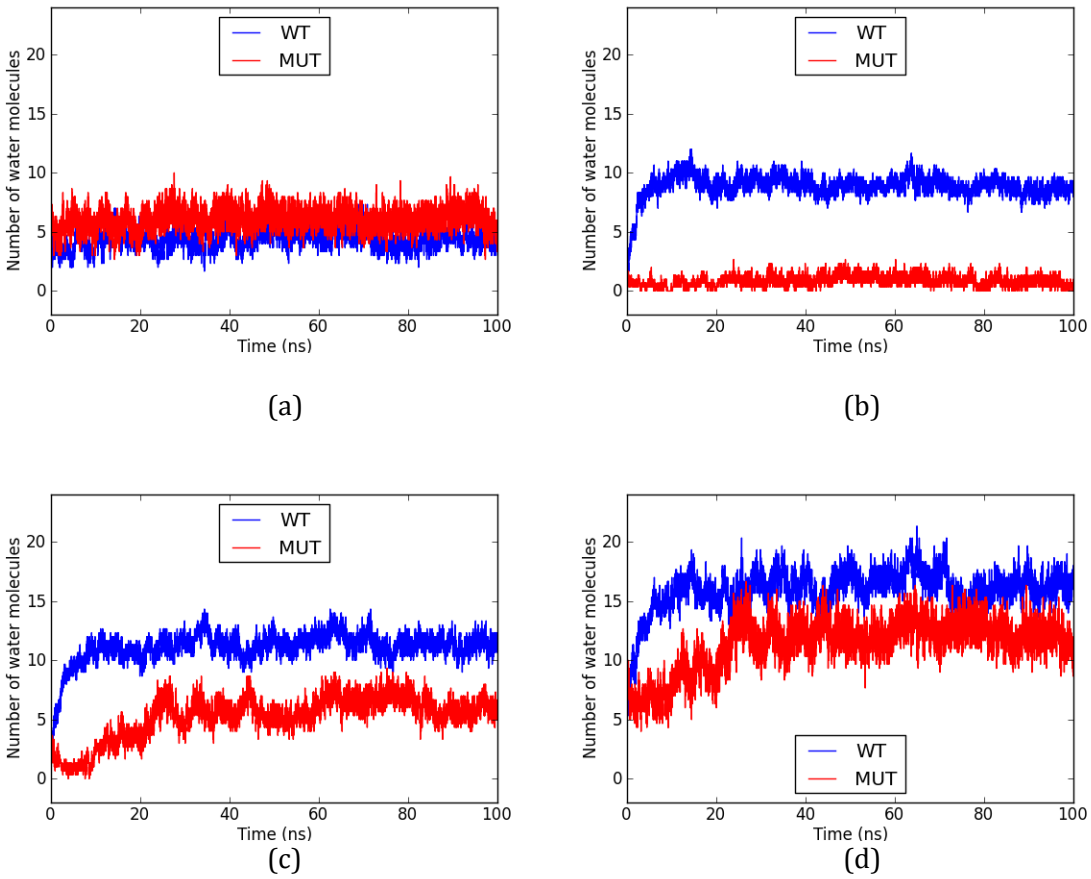


Figure 3.8. The number of water molecules within Ring 2 (a), Ring 3 (b), Ring 4(c), and Ring 2+ Ring 3+ Ring 4 (d). Here the data were the average of three trajectories. WT is referred to wildtype, and MUT is referred to Mutant C.

3.3.3.3. Vacuum and protein stability

An accompanying experimental study, unfortunately, found that the Cel48A mutants were not stable. The failure in protein folding of the mutants might partially result from steric clashes by the heavily concentrated Phe residues in the protein. In addition, the hydrophobic effect of Cel48A water pore mutants was concomitant with cavity creation within the pore region. Pratt and Chandler's molecular theory on hydrophobic effects demonstrated that the hydrophobic effect is important to the assembly of protein into functional complexes, and small cavities induce a free energy cost [140, 159].

The free energy cost for the formation of a small cavity in the water pore of Mutant C was estimated. This free energy cost for generating small cavities in water can be estimated as [160]:

$$\Delta G \approx k_B T \frac{\rho^2 V^2}{2\chi_V} + k_B T(\ln 2\pi\chi_V)/2$$

where V is the volume of the cavity, ρ is the number density of the solvent (water), and χ_V is the mean-square fluctuation in the number of molecules that occupy the cavity volume in the pure liquid. χ_V can be calculated as:

$$\chi_V = \langle(\delta N)^2\rangle_V = \rho V + \rho^2 \int_V d\mathbf{r} \int_V d\mathbf{r}'[g(|\mathbf{r}-\mathbf{r}'|) - 1]$$

where $g(r)$ is the radial distribution function of pure water oxygen-oxygen pairs. Here we defined the cavity volume as the vacuum space in the pore that could accommodate a given number of water molecules. That is, the cavity volume was equal to the number of water molecules that could fit into the vacuum multiplied by the volume of an individual water molecule in the pure liquid at the desired temperature. The total volume of the cavity was approximated to be a sphere. Hence, χ_V was derived to be:

$$\chi_V = \rho V + \frac{16\pi^2}{3}\rho^2 {r_0}^3 \int_0^{r_0} dn\,[n^2\, g(n)]$$

where

$$r_0 = \left(\frac{3V}{4\pi}\right)^{\frac{1}{3}}$$

The results of free energy cost corresponding to N, the number of water molecules that could fit in the pore vacuum, is shown in Table 3.6. As was observed in the CelF mutant, 5 water molecules were depleted from the pore, giving rise to a free energy cost of ~2.0 kcal/mol, whereas in Cel48A Mutant C, the vacuum volume was equivalent to 1 or 2 water molecules, corresponding to 1 ~ 1.5 kcal/mol of free energy cost. Because the free energy change for protein unfolding is on the order of 5 kcal/mol, this result indicated that the mutants might possibly be unstable in an aqueous environment. Additionally, the interacting Phe-Phe pair in vacuum or in water favors stacked

arrangement, and the minimum binding energy for Phe-Phe pair in such arrangement is about -4.4 kcal/mol, according the potential of mean force calculation [161]. As the Phe residues in the water pore structure might not be able to form such favorable interaction, they might cause instability of the protein and further lead to protein misfolding.

Table 3.6. Free energy cost of small cavity creation in the water pore

N	ΔG (kcal/mol)
1	1.092
2	1.490
3	1.743
4	1.915
5	2.037

Note: N is the number of water molecules that the vacuum accommodates.

3.3.3.4. The pathways of the hydrolytic water molecules

While the simulations demonstrated that the water molecules could easily move through the pore from the protein surface to the active site in the wildtype protein (Figure 3.9), supporting the hypothesis that this pore has possible mechanistic significance, the Cel48A wild type simulations also offered evidence of other hydrolytic water pathways as well (Figure 3.10). Analysis of the hydrolytic water pathway demonstrated that the water molecules could also come from the active site tunnel exit, or work their way through the bottom loops of the protein (Table 3.7). While in Mutant C water could not come from the water pore, the water molecules at the active site exchanged much more frequently than that of the wildtype.

Table 3.7. Pathways of water molecules that have occurred at the active site.

	The number of water molecules					
	WT_rep1	WT_rep2	WT_rep3	CMUT_rep1	CMUT_rep2	CMUT_rep3
Ntot	57	53	49	127	134	250
Path 1	17	5	4	0	41	24
Path 2	26	32	31	122	86	217
Path 3	10	13	12	N/A	N/A	N/A
Path 4	4	3	2	5	7	9

Note:
Ntot: the total number of the selected water molecules
Path 1: Bottom loops → Active site region
Path 2: Tunnel exit → Active site region
Path 3: Water pore → Active site region
Path 4: Active site region → aqueous environment

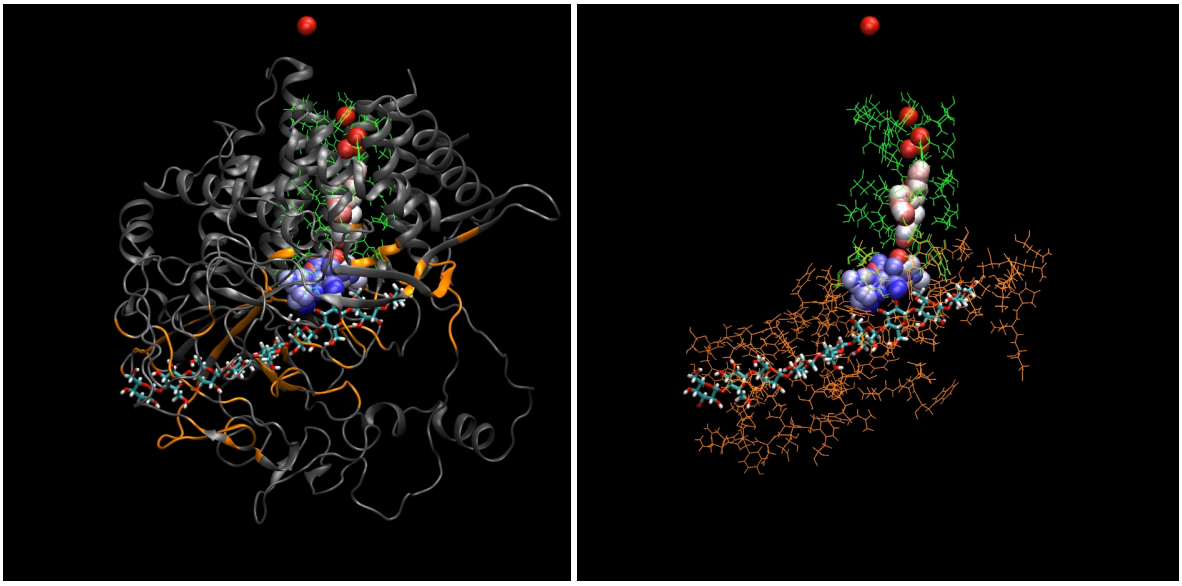
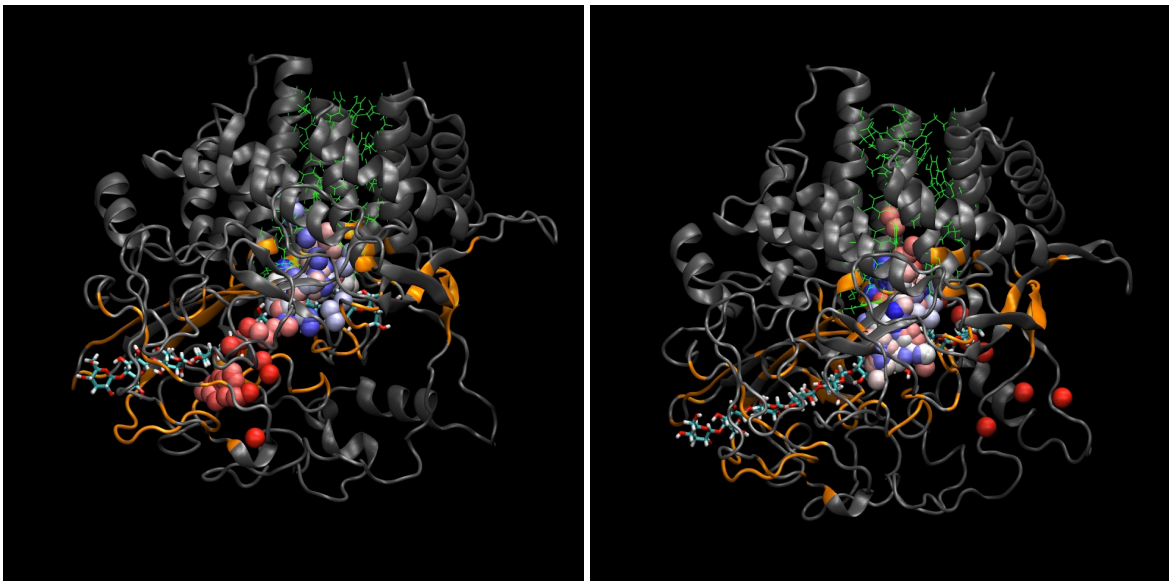


Figure 3.9. A representative trajectory of a single water molecule moving through the water pore of Cel48A wildtype. The water pore residues are colored in green, and the active site tunnel residues are colored in orange. The substrate is shown in licorice and colored by atom type. In these trajectories, only the water oxygen atoms are shown, color coded to represent the time evolution, with the color trend from red to white to blue equivalent to progression from beginning to end.



(a) (b)

Figure 3.10. A representative trajectory of a single water molecule moving from the bottom loops to the active site region (a), and moving from the active site tunnel exit to the active site region (b). The water pore residues are colored in green, and the active site tunnel residues are colored in orange. The substrate is presented in licorice and colored by atom type. In these trajectories, only the water oxygen atoms are shown, color coded to represent the time evolution, with the color trend from red to white to blue equivalent to progression from beginning to end.

3.4. Conclusions

In this study, a unique water-filled pore structure was identified in family 48 cellulases that connect the protein surface with the active site. Therefore, it was hypothesized that this pore structure might be of mechanistic importance by providing a convenient pathway to transport hydrolytic water for substrate hydrolysis. Using hydrophobic effects, the water pore mutants of CelF and Cel48A, two representatives of family 48 cellulases, were successfully designed that could block water transport through the pore. Molecular modeling and simulations of the enzymes and their mutants on short time scales showed that the water pore structure in family 48 cellulases provided an easy path for hydrolytic water transport, though not the exclusive one. The hypothesis, though circumstantially plausible, was neither proved nor disproved by the experimental studies (not presented here), due to the inability of the rationally designed mutants to fold properly. The failure in protein folding of the mutants might result from steric clashes of the highly concentrated

Phe's and the unfavorable interactions of the Phe's in the pore structure, as well as the free energy penalty of the small vacuum generated within the inner section of the pore.

CHAPTER 4

STUDYING THE β-D-GLUCOPYRANOSE BINDING AFFINITY ON CELF SURFACE

4.1. Introduction

In nature, cellulose consists of crystalline regions combined with amorphous regions in varying amounts, depending on the source. To allow for cellulose hydrolysis by cellulases, recognition of cellulose needs to be achieved by the cellulases as an initial step. Many cellulase-producing microorganisms produce enzymes with non-catalytic carbohydrate-binding domains (CBMs) in addition to the catalytic domains (CDs), promoting the association of the enzymes with the substrates [162]. Moreover, the crystal structures of many CDs, particularly those of processive exocellulases such as *Trichoderma reesei* Cel6A [163] and Cel7A [164], and family 48 exocellulases from various microorganisms [144, 165], illustrate that Trp residues are present at the entrances and exits of the active site tunnels. It was suggested that the Trp residues at the tunnel entrances might function in cellulose recognition and acquisition, and those at the tunnel exits might function in stabilizing the product side of the substrate prior to its hydrolysis [166]. Free energy calculations using molecular dynamics simulations have provided support to such mechanisms at molecular level. For example, Payne and coworkers calculated the relative ligand binding free energies between the wildtype and the Ala mutants that corresponded to each of the four tunnel Trp residues in *T. reesei* Cel6A, and revealed that removing only the Trp's at the tunnel entrance and exit could dramatically impact the binding affinity to the substrate [166]. Wohlert and coworkers calculated the free energy landscape for the interaction between indole (the sidechain of Trp) and β-D-glucopyranose (the unit structure of cellulose) in aqueous solution, showing that the two species favored stacking interaction, corresponding to a binding energy of ~1.2 KJ/mol [167].

Family 48 cellulases have a long active site tunnel (~43Å), which contains four conserved Trp residues that form stacking interactions with the unit structure of the cellooligomer substrate

in the crystal structures, in which two are located at the entrance, one at the exit, and the other near the active site (Figure 4.1) [144, 165]. The active site tunnel guides the cellulose chain to the active site. Similarly to the previously introduced studies, we hypothesized that the Trp's at the tunnel entrance and exit had relatively higher affinity for the substrate than the protein surface on average. In this study, β-D-glucopyranose was used to represent the cellulose unit structure, and the binding pattern of glucoses on the surface of CelF, a family 48 exocellulase from *Clostridium cellulolyticum* was characterized. Because glucose is an osmolyte that favors aqueous solution rather than tending to bind on protein surfaces, a weak but sufficiently significant binding affinity between glucose and the tunnel entrance and exit of the globular protein, would support the previously suggested cellulose binding mechanism of the processive exocellulases.

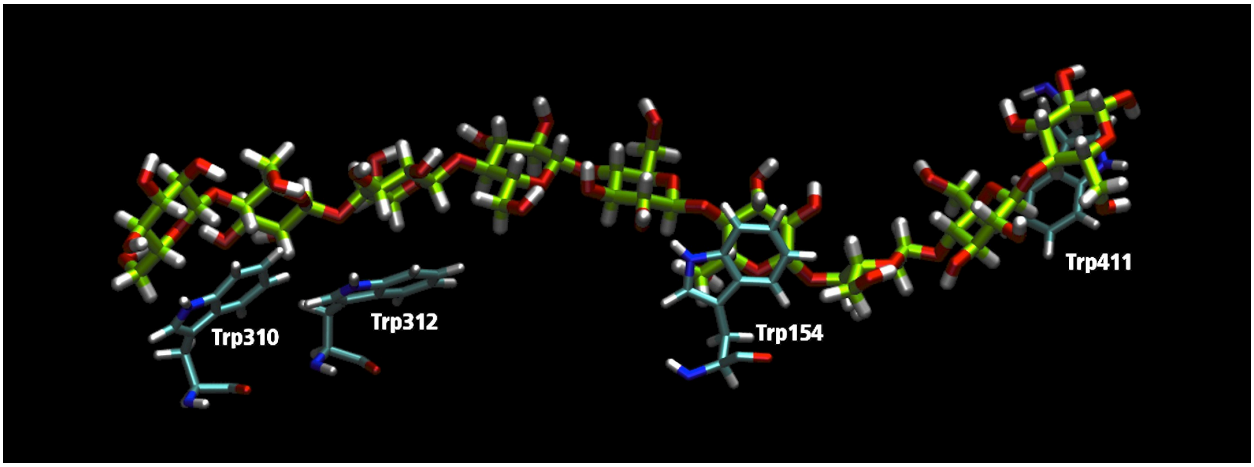


Figure 4.1. Four Trp's in the active site tunnel of CelF, a representative of family 48 cellulases, stack onto the unit structure of the cellooligomer chain. The left side is the tunnel entrance and the right side is the tunnel exit.

4.2. Methods

4.2.1. Structural preparation

The crystal structure of CelF was used as the starting structure in the simulation. Several CelF crystal structures have been solved with various ligands occupying the substrate and product

sites in the protein [143, 144, 168]. These crystal structures have high levels of similarity. In particular, the most recently solved CelF structures correspond to two CelF single mutants, presenting two substrate pathways [144]. The one with a "lower" substrate pathway (E55Q) was considered a more realistic model to represent wildtype CelF, and was used to recover the structure of wildtype CelF.

The molecular system was constructed by positioning a CelF molecule in the center of a large cubic box with the dimension of ~96Å containing a concentrated solution of β-D-glucopyranose. This box contained 864 glucose molecules in total, and was built by assembling 27 small glucose cubes together, each of which contained 32 glucose molecules that were randomly generated with no atomic overlap. Next, water molecules were added to fill in the vacant space in the large cubic system, and sodium counter ions were placed randomly to neutralize the system. Overall, the system contained 1 CelF, 707 β-glucoses, 13 sodium cations, and 20,787 water molecules. It thus represented a glucose concentration of ~1.89 molal.

4.2.2. Molecular simulations

Molecular dynamics (MD) simulation was used to simulate the molecular system. The CHARMM22/CMAP force field [103, 107], the all sugar carbohydrate force field [108], and the TIP3P model [110] were used to describe the protein, the cellooligomer, and the water molecules. The MD program CHARMM [111, 112] was used to build the molecular system. The CHAMBER program [114] was used to convert the CHARMM files into AMBER format. The PMEMD engine of AMBER [113] was used to carry out the MD simulations. The SHAKE algorithm [169] was applied in the simulation to constrain the bond distances involving hydrogen atoms. The nonbond cutoff distance was 8Å.

To prepare the system, the solvent of the system was subjected to minimization, including 100 steps of steepest descent followed by 50 steps of conjugate gradient. Then, the solvent was

thermalized at constant volume from 0 K to 300 K for 20 ps. The production run of the molecular dynamics simulation was carried out under constant temperature at 300 K and constant pressure at 1 atm for 100 ns with a timestep of 2 fs. Constant temperature coupling was regulated by a Langevin thermostat [125-127], and constant pressure coupling was controlled by the Berendsen weak coupling algorithm [120].

4.2.3. Volume density map calculation

Every 10 ps of the production run trajectory was extracted and saved into a new sampled trajectory, which contained 10000 frames representing the entire 100 ns of the production run. The trajectory of the production run generated by AMBER was initially not wrapped into the primitive box of the system. The post-trajectory processing tool Ptraj of AMBER Tool [113] allows for wrapping the trajectory into the primitive cubic box with respect to specified atom selections. Using this tool, the new trajectory was wrapped into a cubic box centering on CelF with a RMS fitting algorithm, which superimposed the CelF in each frame onto that in the initial frame while moving all the other molecules inside the cubic box. All frames of this recentered trajectory were used to calculate the volume density map of glucose.

The volume density map was calculated using the VolMap tool of VMD. The selected atoms, which were the glucose ring heavy atoms (C1, C2, C3, C4, C5, and O5), were used to represent the movement of the glucose molecules. The resolution of the density map was set to be 1Å, and the weight was set to be the "occupancy", with 1 referring to the grid occupied and 0 referring to the grid non-occupied.

4.3. Results and discussions

4.3.1. Volume density map

The volume density map of β-D-glucopyranose at the isovalue of 0.0185 (unit: the number of atoms per $Å^3$) illustrates that multiple small regions on the CelF protein surface have a high tendency to bind to glucose (Figure 4.2). In particular, the density clouds 1, 2, and 3 have large volumes that are comparable to the size of a glucose molecule, indicating a relatively stable binding conformation. These three density clouds located at the active site tunnel entrance -6 subsite, -7 subsite, and at the tunnel exit around the +1 and +2 subsites, respectively, stacking onto a Trp residue at each site, suggesting that Trp residues facilitate cellulose chain binding to the tunnel. This result is consistent with the theoretical study by Wohlert and coworkers, which demonstrated weak binding affinity between indole and β-D-glucopyranose via a stacking interaction [167].

Insufficient sampling remains an issue for MD simulations on large molecular systems like this. Among the three major density clouds, only cloud 2 is formed by relatively frequent glucose exchanges, with 17 exchanges over the 100 ns simulation (Table 4.1). Cloud 1 locates in the inner region of the tunnel entrance, and the presence of glucose in the cloud 2 region can easily prevent glucose from diffusing to the cloud 1 region, causing few glucose exchanges at this site. On the other side, cloud 3 is surrounded by a funnel-shaped tunnel exit. The shape of cloud 3 illustrates a stacking interaction of glucose with Trp411, and the residence time of the 3 glucose molecules at this site is relatively longer, on the order of ~25 ns on average, showing a strong tendency for glucose to bind at this region. This indicates that the system has a lower free energy when glucose molecules bind to the tunnel entrance and exit pockets instead of staying in the aqueous solution. The glucose binding at the tunnel exit also indicates product inhibition. In addition, the localized glucose molecules in clouds 1, 2, and 3 do not exhibit preferences in terms of binding direction within this short simulation. As family 48 cellulases are known to acquire cellulose chains from their reducing end, other mechanisms might exist for progressing of the correct chain end in the tunnel.

Table 4.1. The number of glucose exchanges in each density cloud

Cloud label	Number of glucose exchanges*
1	2
2	17
3	3
4	42
5	10
6	22
7	48
8	31
9	55
10	37
11	28
12	55
13	58
14	39
15	46
16	29

*: This is the count of localized glucose molecules with different residue IDs.

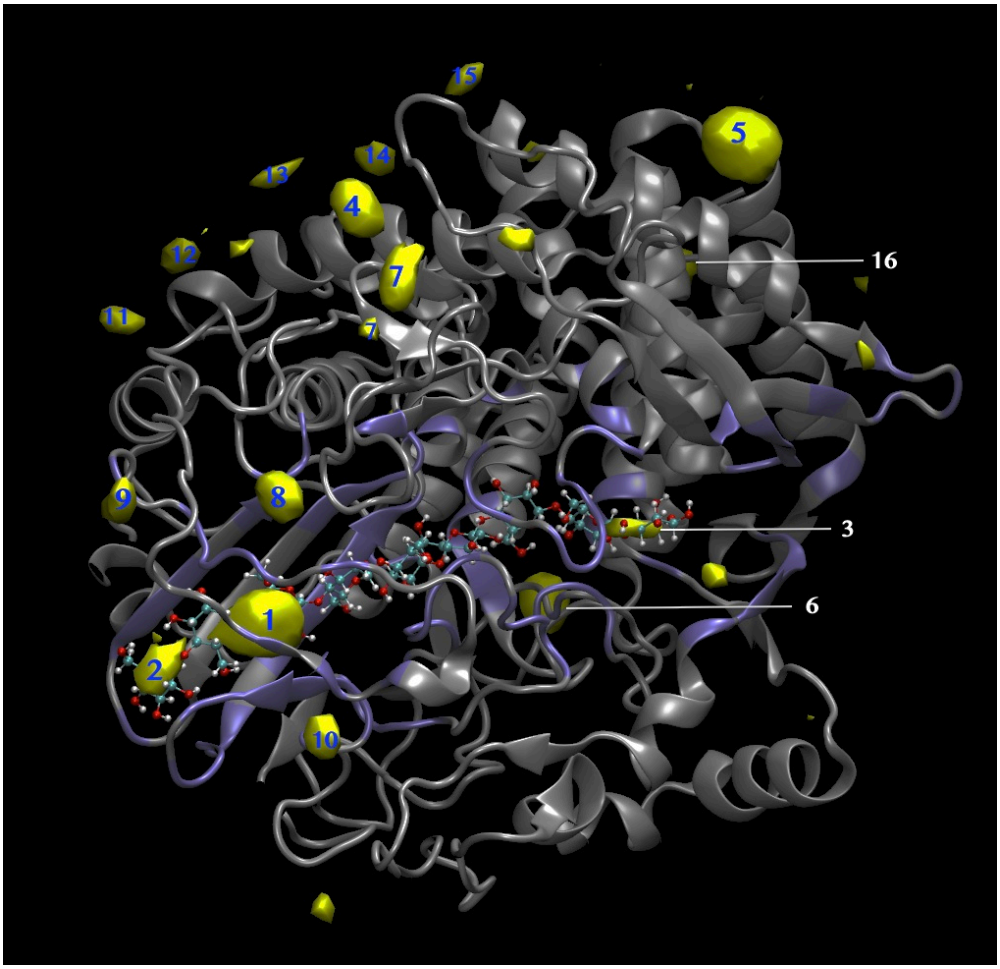


Figure 4.2. The volume density map of β-D-glucopyranose ring heavy atoms (C1, C2, C3, C4, C5, and O5) at the isovalue of 0.0185 (unit: the number of atoms per Å^3). The CelF backbone is shown in "NewCartoon" representation with the active site tunnel residues highlighted in purple. The density clouds of the atom selections are shown in yellow. The cellooligomer (DP=9) in the active site tunnel is superimposed to highlight the tunnel, and it is not present in the MD simulation.

4.3.2. Hydrogen bond characterization

Hydrogen bonding between glucose molecules and the local amino acid residues certainly played an important role in the glucose localization. We analyzed the temporal evolution of the intermolecular hydrogen bonds between the localized glucose molecules in each density cloud and the local protein residues (Figure 4.3b-4.18b). Geometric criteria for the hydrogen bonds were used; these were: a donor-acceptor distance less than 3.4Å and an angle cutoff greater than 145^o. Additionally, the occurrence of glucose localization in each density cloud was characterized; the

occurrence was set to be 1 if any atom of a glucose molecule was within 1Å of the geometric center of the density cloud, whereas the occurrence was set to be 0 otherwise (Figure 4.3b-4.18b). The results demonstrated that formation of multiple hydrogen bonds was a concomitant of the occurrence of the localized glucose molecules in each density cloud, and at least 4 instantaneous hydrogen bonds were required for glucose binding to the local protein surface. The amino acid residues around each density cloud are displayed in Figure 4.3a-4.18a.

4.3.3. Binding free energy

The binding free energy for glucose molecules in the density cloud 2 was estimated from the volume density map using the same procedures previously used for other systems [139, 170]. This can be achieved by calculating a host–guest type equilibrium constant from the concentration of bound glucose:

$$K_{eq} = \frac{[glucose \cdot protein]}{[glucose][protein]}$$

where $[glucose \cdot protein]$, $[glucose]$, and $[protein]$ are respectively the concentrations of bound glucose, the free glucose, and the protein. This can be used to calculate the binding free energy:

$$\Delta G = -RT \ln\left(\frac{K_{eq}}{K^0}\right)$$

where the standard state is defined as the binding site volume occupied by bulk density glucose.

Let $[glucose \cdot protein] = x$ for the system under study, since x is very small,

$$K_{eq} = \frac{x}{(1-x)^2} \approx x$$

Similarly, let $[glucose \cdot protein] = x_0$ for the standard system,

$$K^0 = \frac{x_0}{(1-x_0)^2} \approx x_0$$

Therefore, the binding free energy of weak binding interaction can be estimated as:

$$\Delta G \approx -RT \ln\left(\frac{x}{x_0}\right)$$

The definition of the binding site is arbitrary since the choice of the density cutoff used to specify it is arbitrarily selected. However, the binding site is fairly well localized since the density falls off steeply with displacement away from the center of the ring. Therefore, within a good range of cutoff values, the calculated energy does not change rapidly with the cutoff, allowing a qualitative estimate of the binding affinity that is not strongly dependent on the selection of cutoff. Another uncertainty in this calculation results from the difficulty in estimating bulk density. Here the bulk density of the selected atoms is estimated to be the averaged density of all the grids that have a smaller value than the homogeneous density of the selected atoms, which is 0.00567 (unit: the number of atoms per $Å^3$). The binding free energy of glucose in the density cloud 2 is ~0.8 kcal/mol (Table 4.2), which is within the same magnitude as the calculated Trp-glucose binding energy in melittin [24].

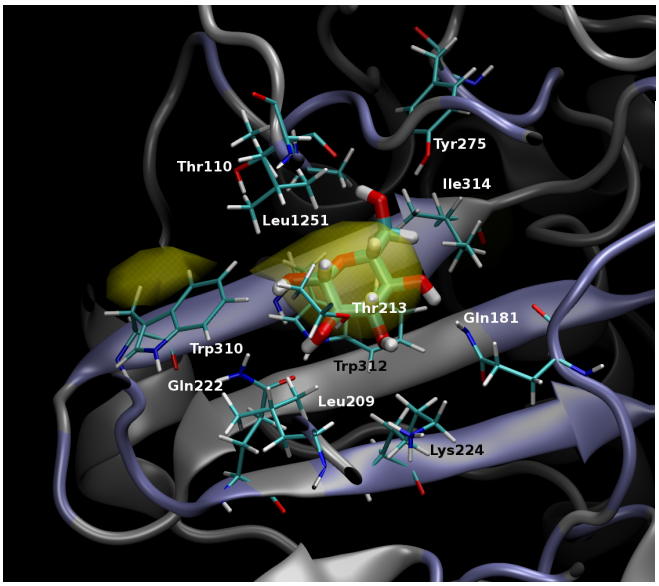
Table 4.2. The binding energy of β-D-glucopyranose at the tunnel entrance.

Contour level (× bulk density)	K_{eq}	Calculated binding energy (kcal/mol)
3	3.770	-0.785
3.5	3.911	-0.807
4	4.247	-0.855

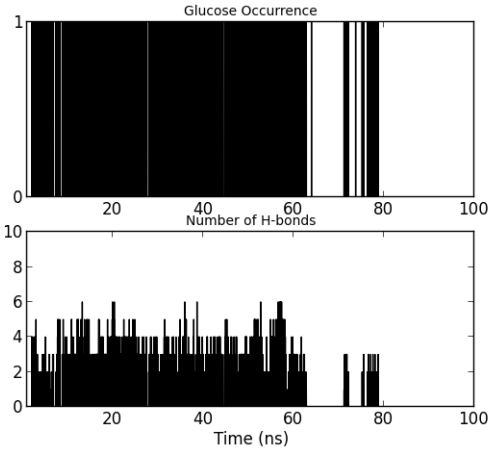
4.4. Conclusions

β-D-glucopyranose, the monomer repeat unit of a cellulose chain, has a weak binding affinity to certain regions on the surface of CelF, which is a family 48 exocellulase, as is shown by the calculated volume density map of glucose on CelF surface. The localized glucose molecules make multiple hydrogen bonds with the local protein residues. In particular, glucose tends to bind at the active site tunnel entrance and exit, where it forms a stacking interaction with Trp residues. From the 100 ns simulation, a nominally sufficient number of glucose exchanges at the tunnel entrance -7 subsite were observed. The binding energy between glucose and this specific site was

estimated to be ~0.8 kcal/mol. This result, to some degree, supports the general assumption that family 48 cellulases can acquire cellulose chains through the entrance of the active site tunnel. Additionally, glucose binding at the tunnel exit +1 and +2 subsites was observed to last for a longer period of time (on the order of ~25 ns for each binding event), though only a few glucose exchanges happened over the short simulation time. This binding phenomenon indicates product inhibition at this site.

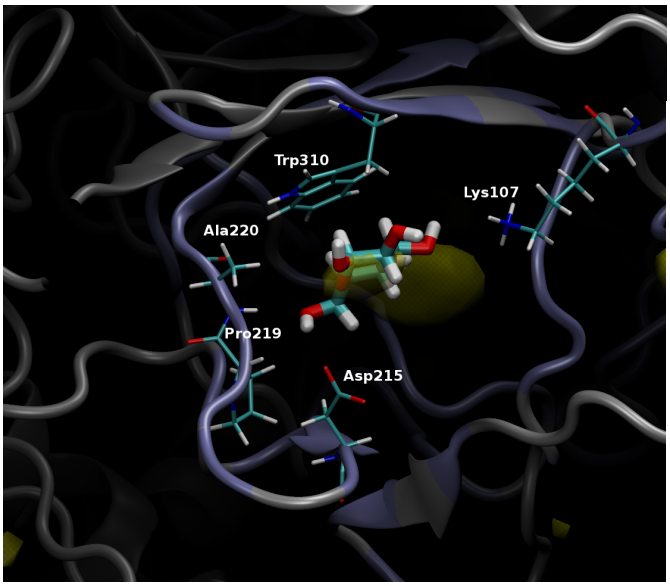


(a)

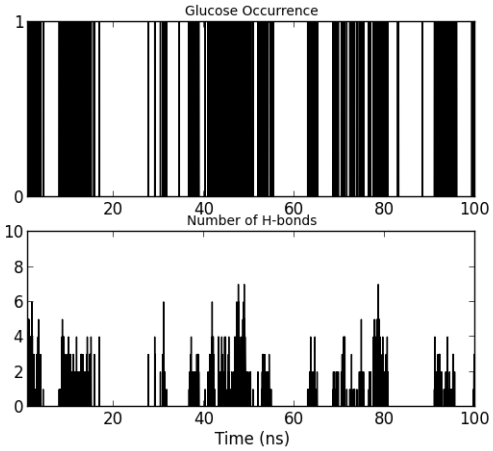


(b)

Figure 4.3. (a) The local protein residues around the density cloud 1; (b) Temporal evolution of the glucose occurrence (upper) and the hydrogen bond (H-bond) number (lower) in the cloud 1.

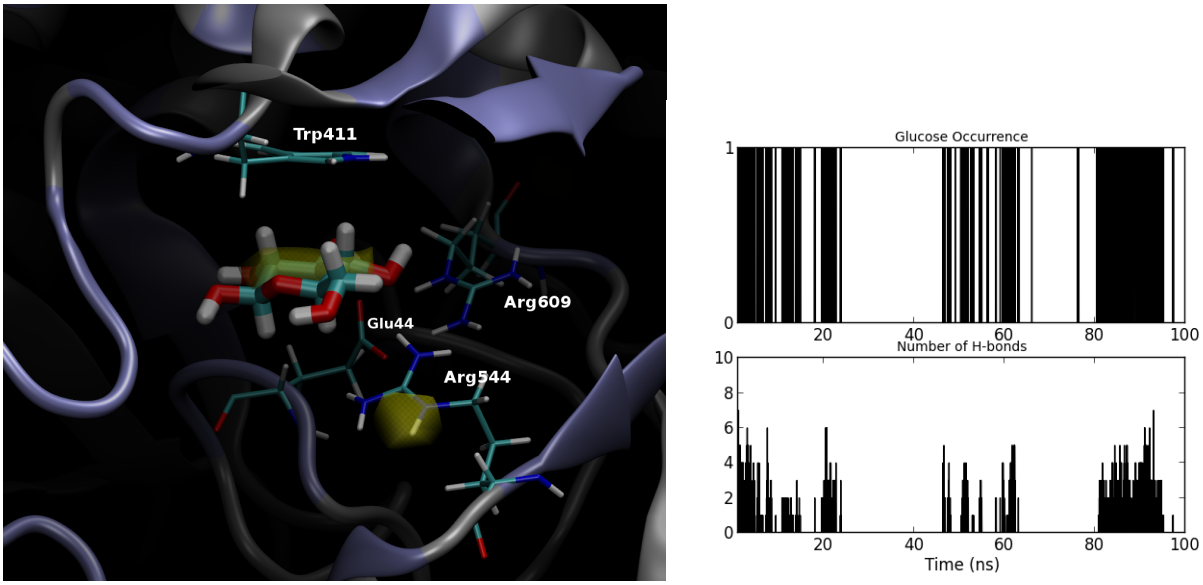


(a)



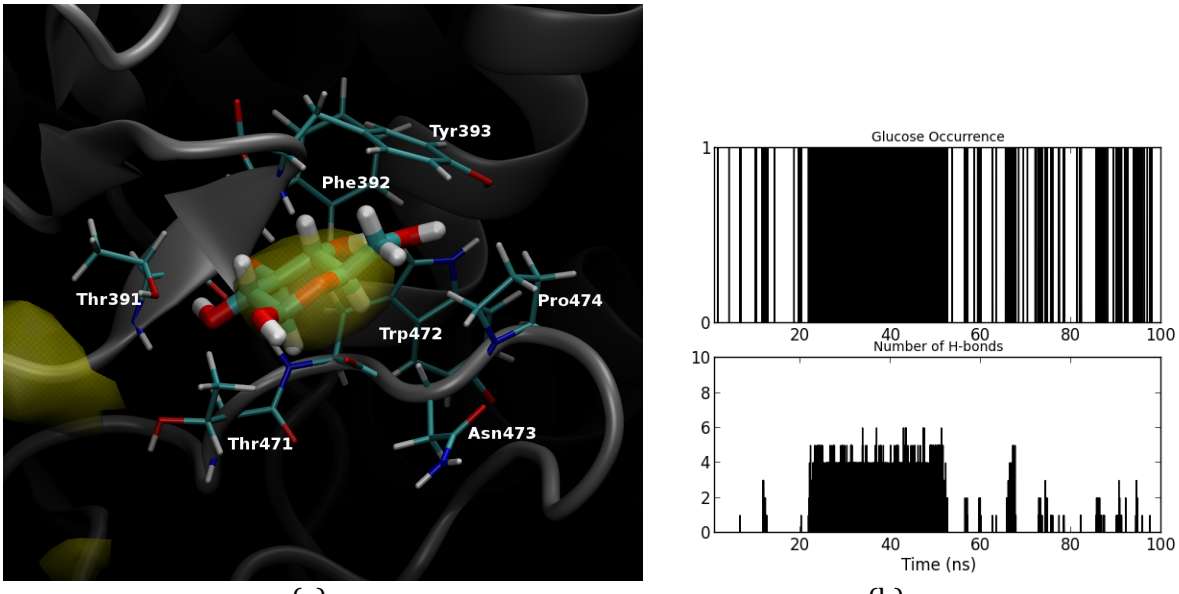
(b)

Figure 4.4. (a) The local protein residues around the density cloud 2; (b) Temporal evolution of the glucose occurrence (upper) and the hydrogen bond (H-bond) number (lower) in the cloud 2.



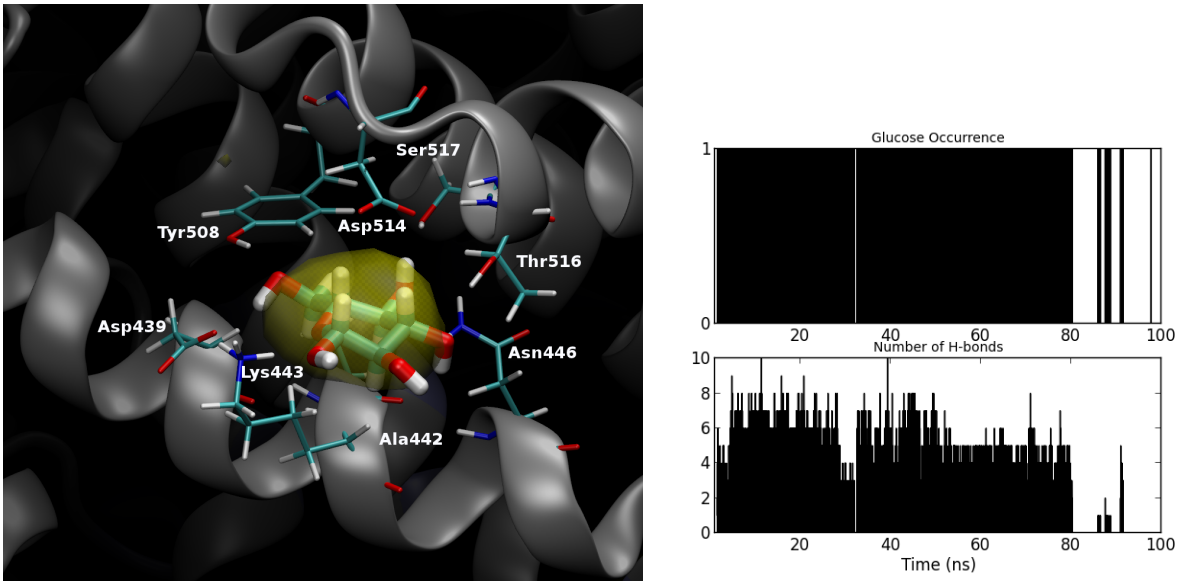
(a) (b)

Figure 4.5. (a) The local protein residues around the density cloud 3; (b) Temporal evolution of the glucose occurrence (upper) and the hydrogen bond (H-bond) number (lower) in the cloud 3.



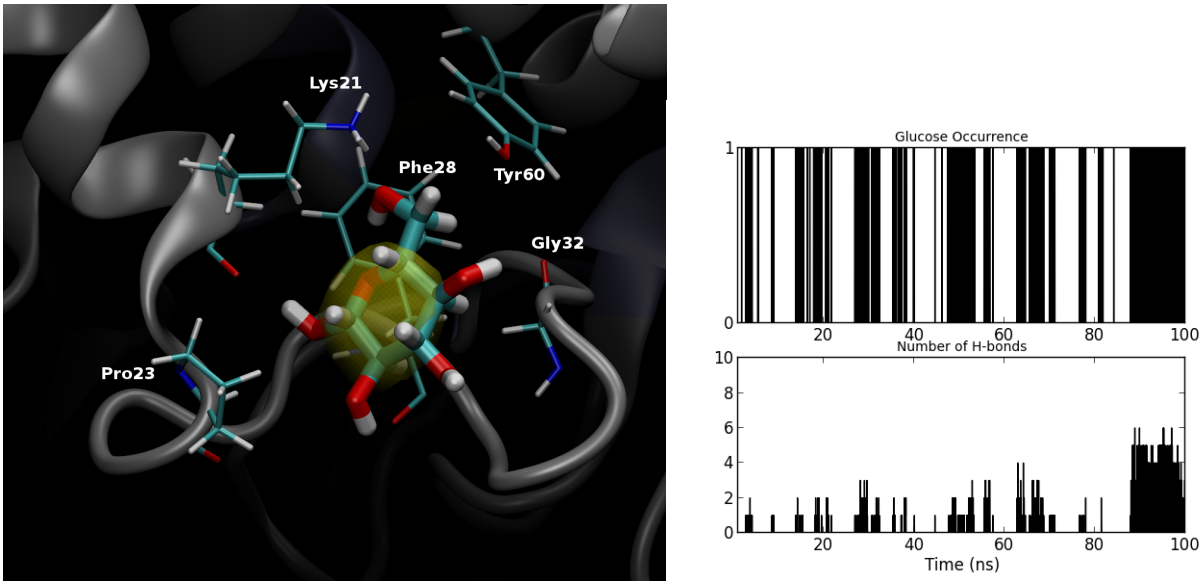
(a) (b)

Figure 4.6. (a) The local protein residues around the density cloud 4; (b) Temporal evolution of the glucose occurrence (upper) and the hydrogen bond (H-bond) number (lower) in the cloud 4.



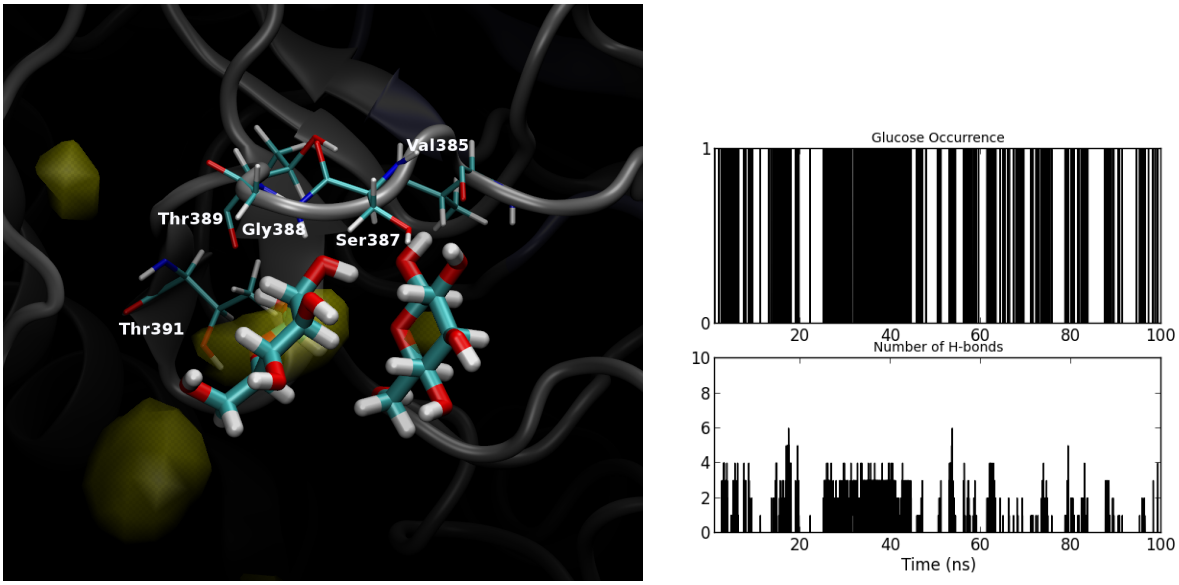
(a) (b)

Figure 4.7. (a) The local protein residues around the density cloud 5; (b) Temporal evolution of the glucose occurrence (upper) and the hydrogen bond (H-bond) number (lower) in the cloud 5.



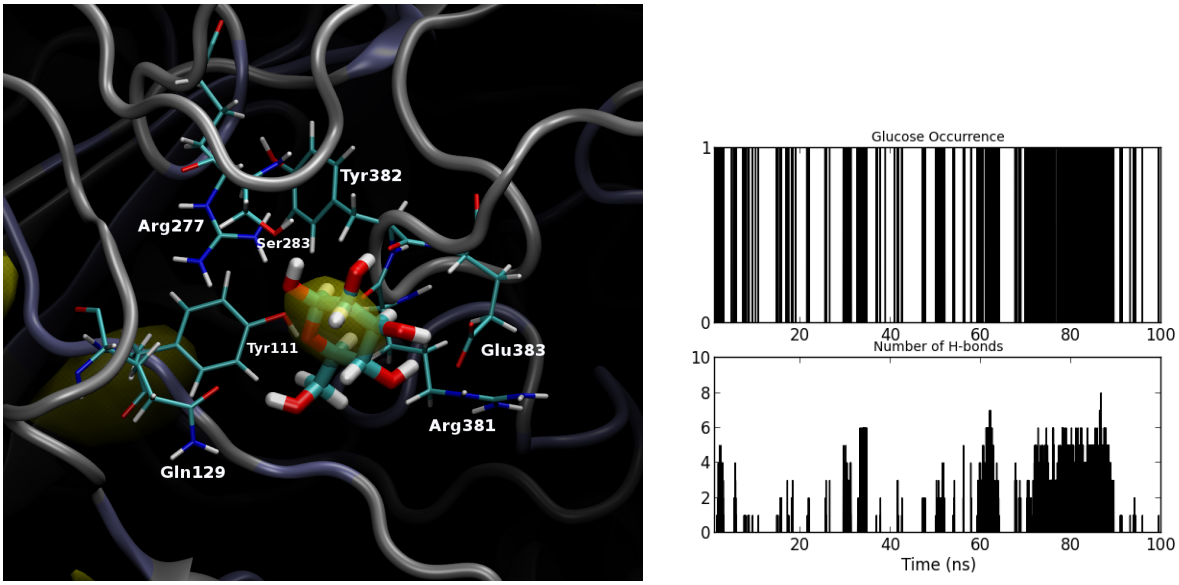
(a) (b)

Figure 4.8. (a) The local protein residues around the density cloud 6; (b) Temporal evolution of the glucose occurrence (upper) and the hydrogen bond (H-bond) number (lower) in the cloud 6.



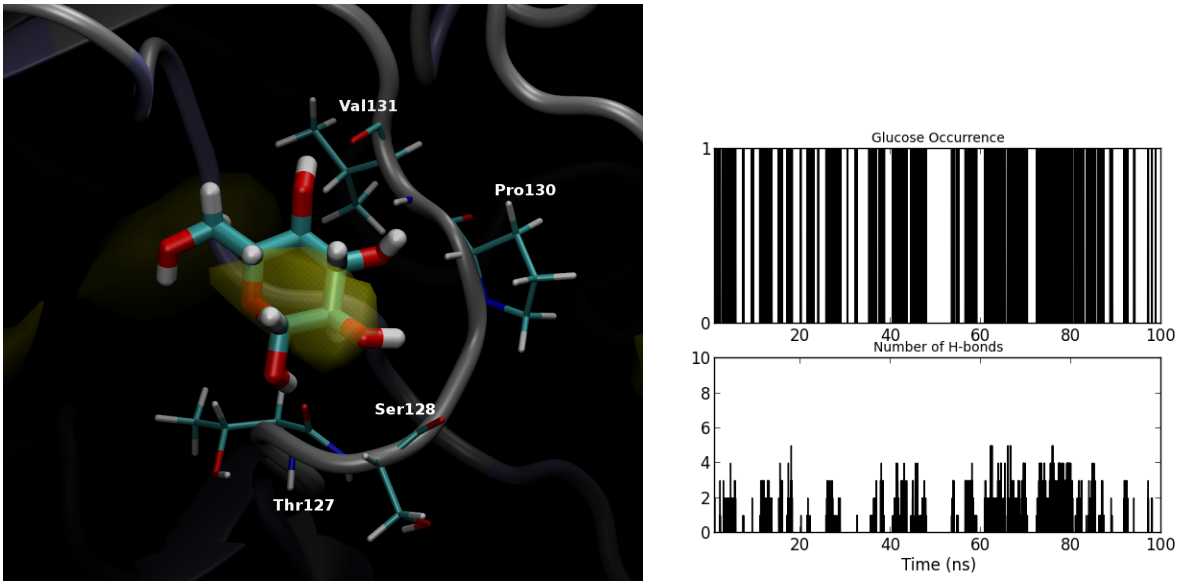
(a) (b)

Figure 4.9. (a) The local protein residues around the density cloud 7; (b) Temporal evolution of the glucose occurrence (upper) and the hydrogen bond (H-bond) number (lower) in the cloud 7.



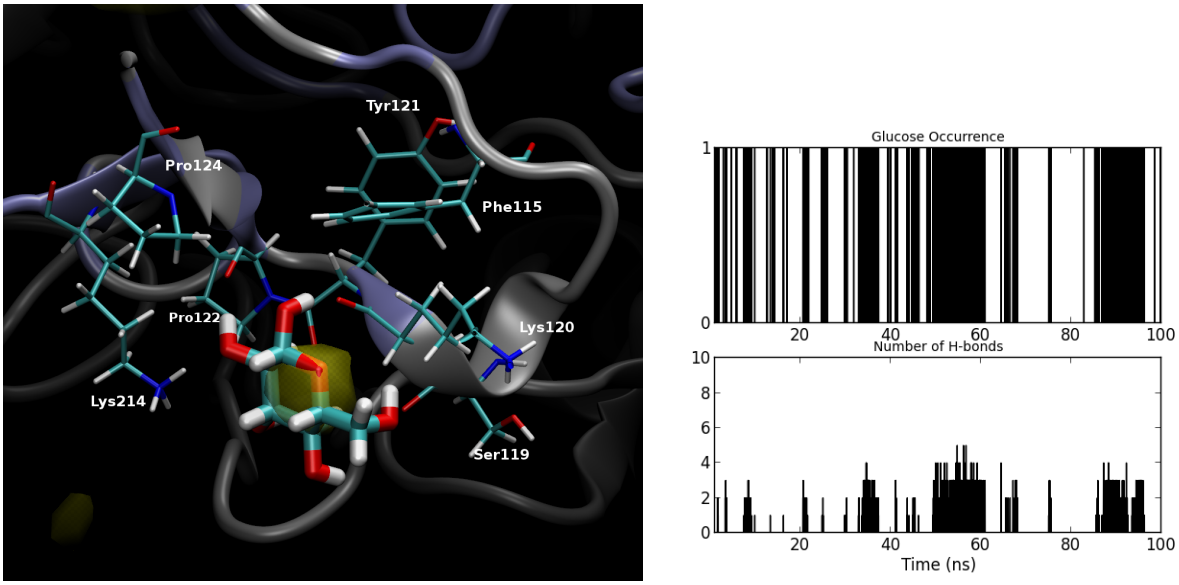
(a) (b)

Figure 4.10. (a) The local protein residues around the density cloud 8; (b) Temporal evolution of the glucose occurrence (upper) and the hydrogen bond (H-bond) number (lower) in the cloud 8.



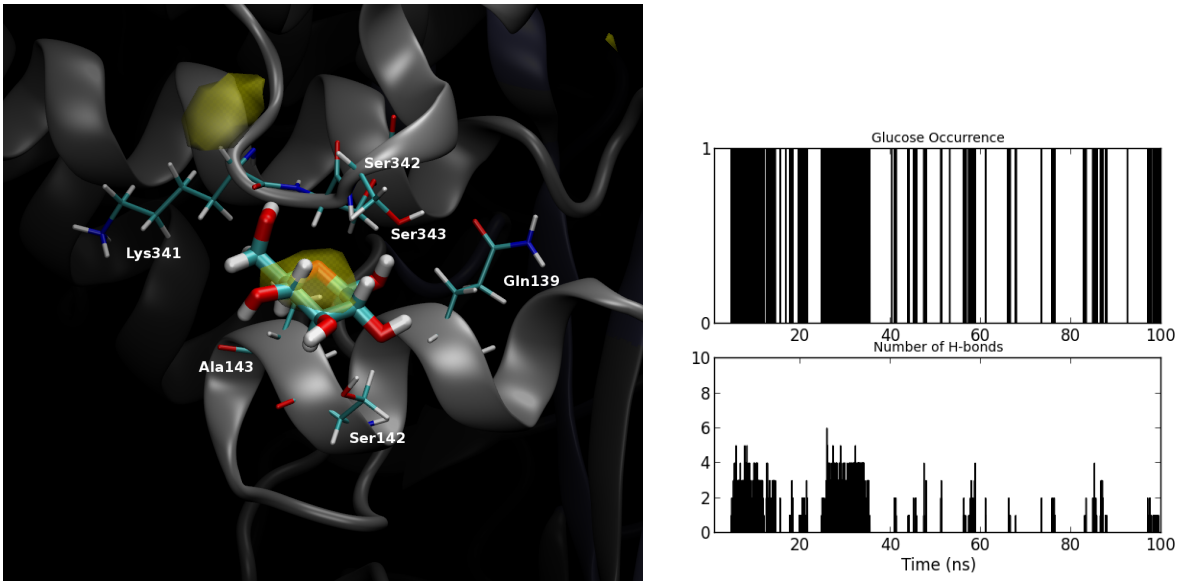
(a) (b)

Figure 4.11. (a) The local protein residues around the density cloud 9; (b) Temporal evolution of the glucose occurrence (upper) and the hydrogen bond (H-bond) number (lower) in the cloud 9.



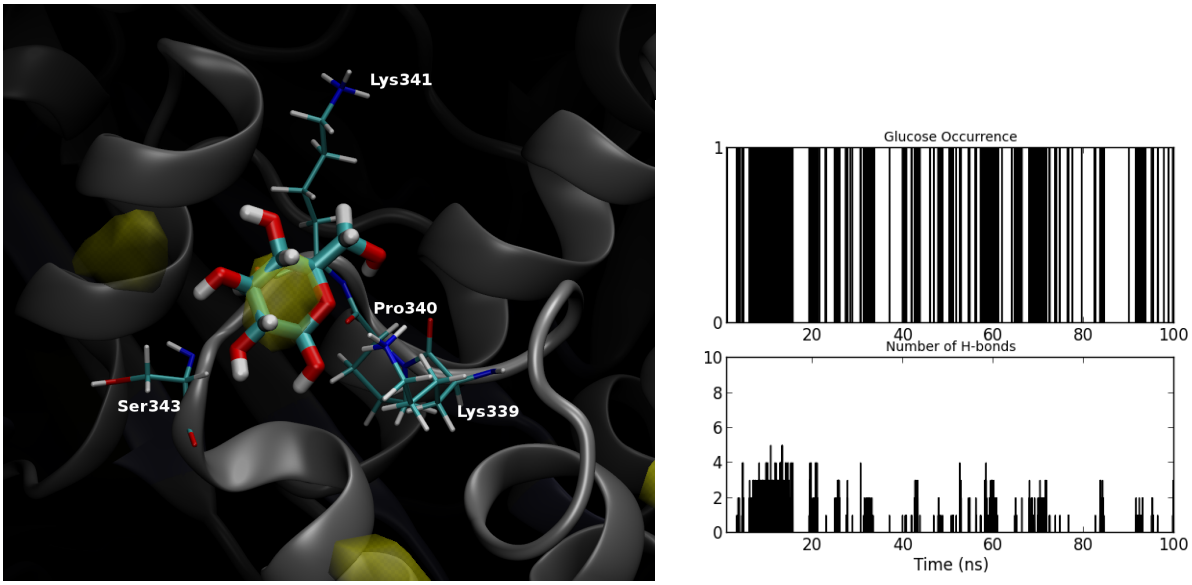
(a) (b)

Figure 4.12. (a) The local protein residues around the density cloud 10; (b) Temporal evolution of the glucose occurrence (upper) and the hydrogen bond (H-bond) number (lower) in the cloud 10.



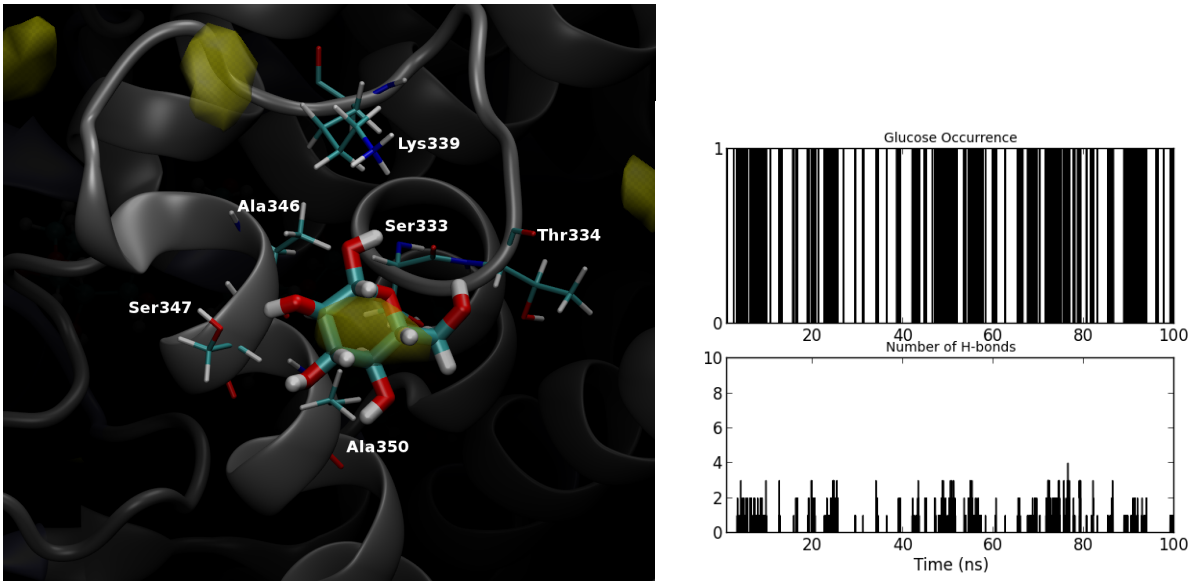
(a) (b)

Figure 4.13. (a) The local protein residues around the density cloud 11; (b) Temporal evolution of the glucose occurrence (upper) and the hydrogen bond (H-bond) number (lower) in the cloud 11.



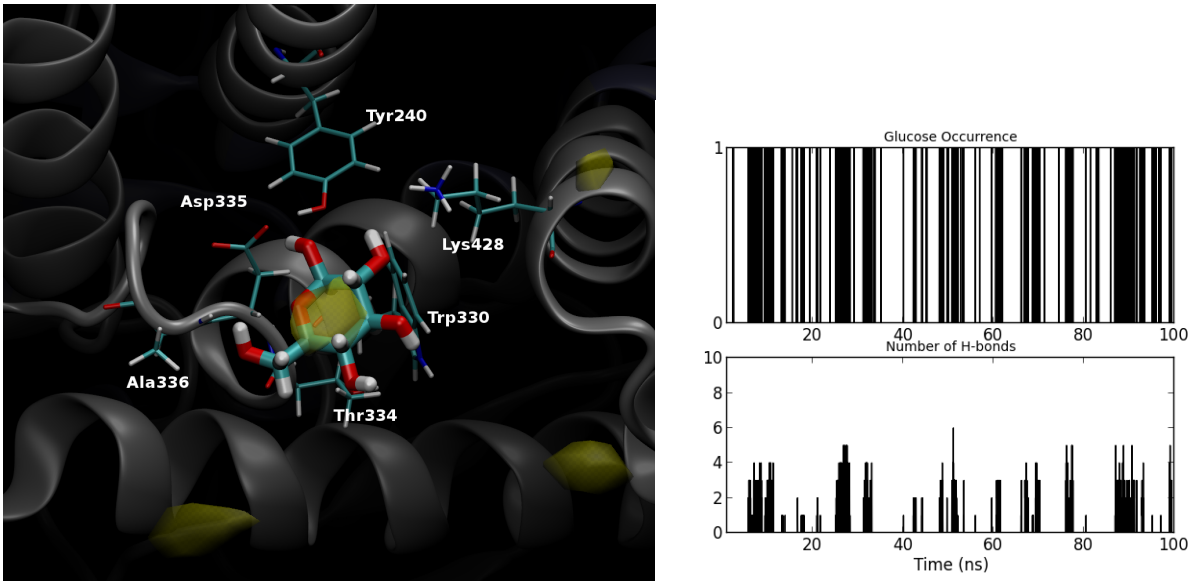
(a) (b)

Figure 4.14. (a) The local protein residues around the density cloud 12; (b) Temporal evolution of the glucose occurrence (upper) and the hydrogen bond (H-bond) number (lower) in the cloud 12.



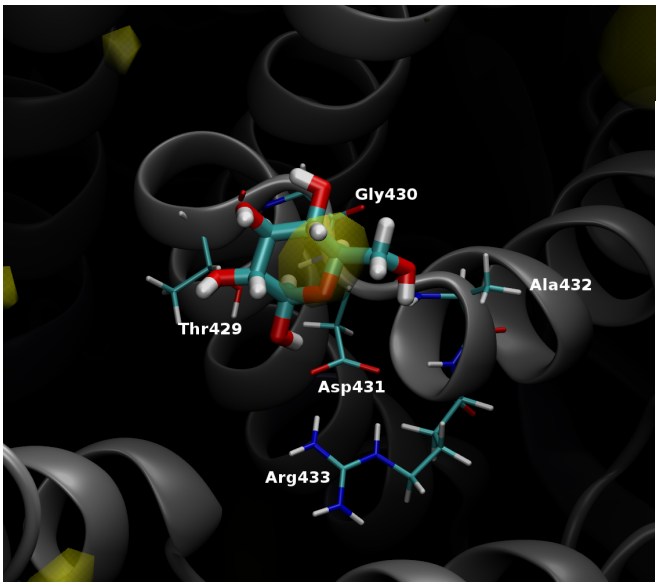
(a) (b)

Figure 4.15. (a) The local protein residues around the density cloud 13; (b) Temporal evolution of the glucose occurrence (upper) and the hydrogen bond (H-bond) number (lower) in the cloud 13.

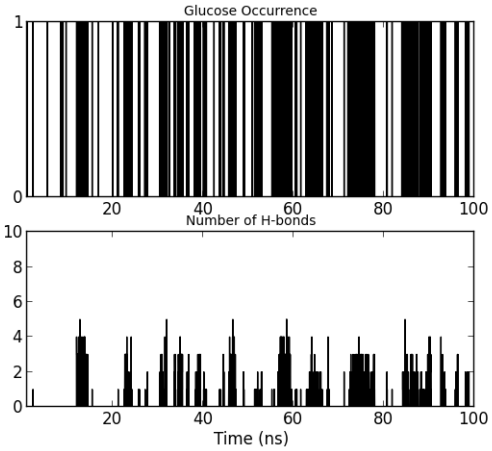


(a) (b)

Figure 4.16. (a) The local protein residues around the density cloud 14; (b) Temporal evolution of the glucose occurrence (upper) and the hydrogen bond (H-bond) number (lower) in the cloud 14.

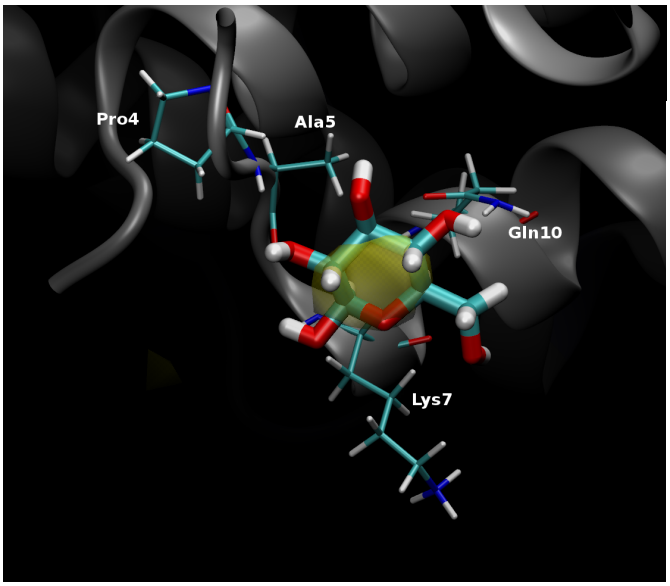


(a)

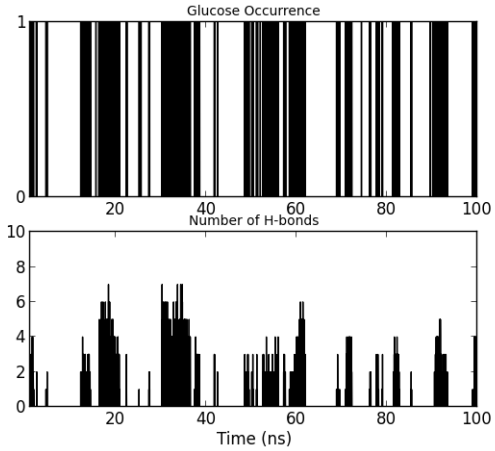


(b)

Figure 4.17. (a) The local protein residues around the density cloud 15; (b) Temporal evolution of the glucose occurrence (upper) and the hydrogen bond (H-bond) number (lower) in the cloud 15.



(a)



(b)

Figure 4.18. (a) The local protein residues around the density cloud 16; (b) Temporal evolution of the glucose occurrence (upper) and the hydrogen bond (H-bond) number (lower) in the cloud 16.

CHAPTER 5

REDUCING LIGAND BINDING FREE ENERGIES IN FAMILY 48 CELLULASES FOR REDUCED LEVELS OF PRODUCT INHIBITION

5.1. Introduction

Lignocellulosic biomass exists abundantly in nature, and is considered a candidate resource for producing biofuel in the form of ethanol. In principle, biomass degradation involves three steps: pretreatment that separates and removes the lignin and hemicellulose components, enzymatic treatment that hydrolyzes cellulose into soluble cellodextrins and glucoses, and fermentation that converts glucoses into ethanol. The biomass conversion process, nevertheless, is very inefficient due to biomass recalcitrance and low enzyme activities. The prevalence of lignin molecules in biomass, wrapped around the cellulose, strongly hinders cellulose degradation [171]. The insolubility of crystalline cellulose in aqueous and most organic solvents also limits its accessibility to the hydrolytic enzymes. Additionally, the catalytic activities of cellulases, which hydrolyze cellulose into soluble cellooligomers (DP≤6), are particularly low. It has been reported that large amounts of enzymes (~25 kg/ton of cellulose) are required to release most of the sugars from biomass at rates compatible with high-throughput processes, and the requirement for such unusually large amounts of enzymes appears to be the single largest cost in the production of cellulosic biofuels [172]. Increasing the catalytic activity of cellulases can potentially reduce the cost of the process and improve the overall biomass conversion rate. To achieve such goals, rational design based on an understanding of the structure-function relationships of various cellulases and their interactions with cellulose substrate can serve as a useful approach.

Family 48 cellobiohydrolases (GH48) are a major group of processive exocellulases that catalyze cellulose hydrolysis from the chain ends and produce mostly cellobiose molecules. More than twenty GH48s have been identified from various microorganisms [32], and among them, the X-

ray crystal structures of four GH48s have been solved, which are: Cel48 from *Bacillus pumilus* (Vladimir V. Lunin, personal communication); CelA from *Caldicellulosiruptor bescii* [173]; CelS from *Clostridium thermocellum* [165]; and CelF from *Clostridium cellulolyticum* [143, 144, 168]. These structures share common features, including an $(\alpha/\alpha)_6$ barrel structure, and a Trp-rich active site tunnel providing seven substrate subsites preceding the hydrolytic cleavage site and two after it at the tunnel exit. In particular, CelA, CelS, and CelF have comparable structures, whereas Cel48 exhibits longer loops, including one at the tunnel exit (Figure 5.1). The melting temperatures for Cel48, CelF, CelS, and CelA are 45 ˚C, 55 ˚C, 65 ˚C, and 85 ˚C, respectively (Yannick J. Bomble, personal communication). Regarding catalytic activity, Cel48 and CelF favor mesophilic conditions, CelS is thermophilic, and CelA is extremely thermophilic, such that it exhibits optimal activity at 75 ^{o}C and sustains high temperatures up to 90 ^{o}C [148]. Under their optimal conditions, CelA has the highest activity, and Cel48 has the lowest activity (Yannick J. Bomble, personal communication). The thermophilic cellulases are particularly interesting in that they can be added directly to cellulosic biomass immediately after pretreatment under high temperature conditions, increasing energy efficiency. In addition, CelS and CelF are critical components of cellulosomes [149], which are complexed assemblies of cellulases and are more efficient in cellulosic biomass degradation than simple combinations of synergistic cellulases.

The active site tunnel provides nine substrate subsites, serving as a substrate pathway for processive action. These subsites are named as subsites -7, -6, -5, -4, -3, -2, -1, +1, and +2 from the substrate's nonreducing end at the tunnel entrance to the reducing end at the tunnel exit. It is generally believed that family 48 cellulases can recognize the cellulose chains by their reducing end, and acquire them into the active site tunnel. Subsequently, the cellulose chain progresses through the tunnel until it is in position for hydrolytic reaction. Family 48 cellulases follow an inverting mechanism [28, 32, 146]. In particular, they use a catalytic acid (glutamic acid) and a catalytic base (aspartic acid) to achieve the hydrolysis of glycosidic bonds in cellulose chains. As a result, usually a

cellobiose product is cleaved off and released to the aqueous environment. Next, the cellulose chain progresses forward in the tunnel by a cellobiose unit so that the catalytic cycle is continued. At some point, the cellulases dissociate from the cellulose substrate, halting the processive hydrolysis. The enzymatic activities of family 48 cellulases are reported to be extremely low. For example, the activities of *T. fusca* Cel48A on swollen cellulose, carboxymethyl cellulose, BMCC, and filter paper are, respectively, 0.405, 0.292, 0.191 and 0.068 μmol CB·min^{-1}·μmol enzyme^{-1} [147]. It has been speculated that the small turnover number of GH48s is due to inefficient acquisition of cellulose by the tunnel entrance, slow processivity of the cellulose substrate in the tunnel, and product inhibition. This study focused on understanding the effect of product inhibition, in order to perhaps provide information to help improve the cellulases' activity.

Several studies have reported that the product cellobiose strongly inhibits the activity of the family 48 cellulases, such as *C. thermocellum* CelS [174-176] and *T. fusca* Cel48A [147]. Almost complete inhibition of the *C. thermocellum* cellulosome, in which CelS was a major component, occurred at a concentration of 2% (w:v) cellobiose [175], much lower than the solubility of cellobiose in water (12% w:v). It was postulated that cellobiose could competitively bind to both the tunnel entrance and the tunnel exit. Since it is desirable for the cellulases to possess sufficient binding affinity at the tunnel entrance for substrate recognition and acquisition, this study focused on the product inhibitory effect at the tunnel exit.

The mechanism of product inhibition at the molecular level has not been fully determined by experimental measurements. As indirect evidence, Zhang and others found that the initial attack of cellulose by *Clostridium phytofermentans* Cel48 generated a perceptible amount of cellotetraose (7% on crystalline cellulose and 4% on amorphous cellulose) and cellotriose (15% on crystalline cellulose and 9% on amorphous cellulose), in addition to the major product cellobiose [80]. It is hypothesized that in the transition state, stabilization of the substrate on the product side is required before initiation of substrate cleavage. Hence, the minor production of cellotriose and

cellotetraose indicates their weak binding to the tunnel exit. In addition, crystal formation of family 48 cellulases suggests that the tunnel exit possesses strong sugar-binding affinity. For example, the presence of cellobiose was required in forming the CelS crystal, and when growing the CelS crystal in cellobiose, the solved crystal structure contained a cellobiose only at the tunnel exit, taking up the subsites +1 and +2 [165]. In CelF crystal structures, the subsites +1, +2, and +3 were identified when using cellotriose and cellotetraose as inhibitors, indicating sufficient sugar-binding potential at the tunnel exit [168]. Additionally, a preliminary study on the distribution of β-D-glucopyranose around the CelF surface (in Chapter 4) demonstrated sugar-binding affinity at the tunnel exit.

Characterization of family 48 crystal structures demonstrated that many residues at the tunnel exit are conserved or partially conserved, including a Trp locating at the subsites +1 and +2 and several charged residues including Arg, Asp, and Glu residing along the tunnel exit. The Trp sidechain stacks onto the two β-glucosyl units at the product side, and might play an important role in stabilizing the substrate in the transition state. It is likely that substituting the Trp at this site can weaken the binding between the product and the tunnel exit. In fact, in an endocellulase Cel5A from *Acidothermus cellulolyticus*, when mutating the product-binding Tyr 245 into Gly, the catalytic rate was increased by 40% and the inhibitor constant, K_i, was increased to more than 1480% [94]. However, mutating the product-binding Trp in the exocellulases might weaken the processivity of the exocellulases as well, which is not desirable for crystalline cellulose degradation. On the other hand, the charged residues can form strong electrostatic interactions with the cellobiose product via hydrogen bonding, hindering its escape into the aqueous environment. To better understand the product inhibition in family 48 cellulases, molecular dynamics simulations and free energy calculation were used to evaluate the product expulsion energies in the four crystallized family 48 cellulases. Rational mutants, aiming at reducing the product inhibitory effect, were also suggested and evaluated.

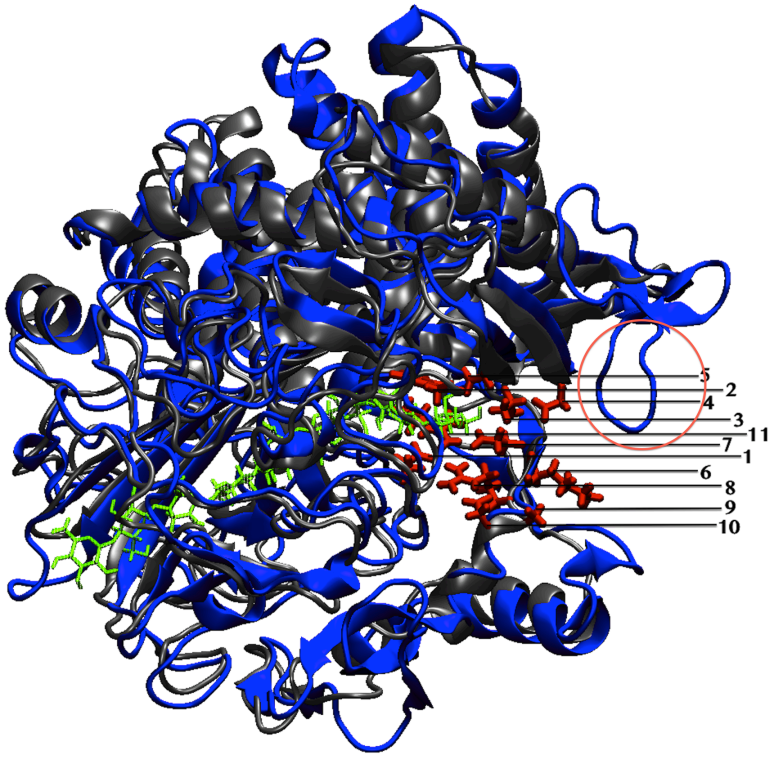


Figure 5.1. Superimposition of the family 48 cellulases crystal structures. CelA CelS, and CelF share very similar structures, and they are shown in grey. Cel48 is shown in blue, and it posses several longer loops compared to the other three structures. In particular, one extra loop in Cel48 locates at the tunnel exit and is highlighted by a red circle. The cellooligomer in the active site tunnel is shown using green sticks. The conserved or partially conserved amino acid residues (see Table 5.1) that form strong interactions with the cellobiose product at the tunnel exit are shown in red sticks, where the labels refer to the Group IDs and the residues correspond to the ones in CelA.

Table 5.1. The featuring residues at the tunnel exit of the four family 48 cellulases

Group ID [1]	CelA	Cel48	CelF	CelS
1	Glu44	Glu38	Glu44	Glu76
2	TRP412	TRP411	TRP411	TRP439
3	Thr463	Gly500	Thr462	Asp490
4	Asp489	Asp530	Asn490	Ser516
5	Asp493	Asp534	Asp494	Asp520
6	Lys547	Glu591	Gln543	Ala577
7	Glu546	Glu590	Glu542	Glu576
8	Arg548	Arg592	Arg544	Arg578
9	Ala549	Glu593	Gly545	Ala579
10	Asp550	Asp594	Asp546	Asp580
11	Arg613	Arg682	Arg609	Arg643

[1]: the residues with the same Group ID are at homologous locations according to protein sequence alignment using the Position-Specific Iterated BLAST method [154].

5.2. Methods

5.2.1. Structure preparations and molecular dynamics simulations

Product inhibition in the four well-characterized family 48 cellulases, particularly CelA, CelF, CelS, and Cel48, was investigated. The atomic models of the cellulases were built based on their X-ray crystal structures. The crystal structures of CelA and Cel48 were provided by the Biosciences Center at the National Renewable Energy Lab. The CelF wildtype structure was converted from the crystal structure of its mutant E55Q with the substrate in the active site tunnel following a lower pathway [144]. The CelS structure was obtained from the reported crystallographic structure [165]. A celloheptaose and a cellobiose from the crystal structures were placed in the active site tunnel taking the positions from subsite -7 to -2 and from subsite +1 to +2 respectively, representing the state immediately after the hydrolysis. Next, each system was put in a truncated octahedral water box. Sodium ions were randomly placed in the water box to neutralize the systems.

The CHARMM22 force field [103] with the CMAP correction [107] was used to describe the protein; the CHARMM36 all-atom carbohydrate force field [108] was used for the cellooligomer, and TIP3P served as the water model [110]. The CHARMM [111, 112] program was used to build the molecular systems. The tool CHAMBER [114] was used to convert the coordinate and structural files and the associated force fields in CHARMM format into AMBER format. The PMEMD engine of AMBER [113] was used to carry out the molecular dynamics simulations, since it gives better performance for the molecular systems compared to CHARMM.

The system preparations prior to production runs included four steps: solvent minimization with 1000 steepest descent steps and 1000 conjugate gradient steps; system minimization with the same strategy; solvent thermalization at constant volume from 0 K to 300 K for 20 ps; and system equilibration in the NPT ensemble at 300 K at 1 atm for 500 ps with a step size of 2 fs. Constant temperature was regulated with a Langevin thermostat, and constant pressure was regulated using

the Berendsen weak coupling algorithm. Subsequently, production runs of the systems were collected in the NVT ensemble at 300 K for 24 ns with a step size of 2 fs. The trajectories of the production runs were used to provide a series of starting structures for the steered molecular dynamics (SMD) simulations. The SMD simulation applies unidirectional forces to the selected atoms to accelerate the movement or conformational changes of the system at a slow but nonzero rate, therefore permitting the study of hypothetical processes within a reasonable time frame. In this study, 40 streams of SMD simulations were conducted for each system, using 40 different starting structures that were extracted from the production run. The criteria for selecting starting structures was that the distance between the C1 atom of the glucose 1 and the C1 atom of glucose 3 was the closest to the most populated distance over the production run. The SMD simulations were performed using the PMEMD engine of AMBER. The reaction coordinate was set to be the distance between the two C1 atoms (Figure 5.2). Over the course of each SMD simulation, external force was exerted on the two C1 atoms, to gradually increase the reaction coordinate at a constant slow speed of 1 Å/ns [177], until a total operation distance of 22 Å was reached, simulating the process of product escape into the aqueous environment after the hydrolysis. The accumulated external work was recorded over the simulations for free energy calculation.

5.2.2. Free energy calculation using a "fast growth" method

The product expulsion energy, in this case the binding free energy between cellobiose and the cellulase, which contains a celloheptaomer bound in the -7 to -1 subsites of the active site tunnel, contributes to the product inhibition. Calculation of protein-ligand binding free energy is intrinsically a very interesting topic. Many methods, for example free energy perturbation, have been successfully implemented in calculating the protein-ligand binding free energy, though such methods for large molecular systems often involve complications at the setup stage [177].

In this study, a nonequilibrium "fast growth" method, also known as Jarzynski's equality, was used to investigate the product expulsion energy. The product expulsion energy was defined to be the free energy difference between the initial state (immediately after the cellobiose was cleaved off the cellulose chain) and the final state (when the cellobiose was released into the aqueous environment). The "fast growth" method refers to an irreversible process, different from the established "slow growth" method, during which the system is driven reversibly from one state to the other [178].

Jarzynski's equality states that the free energy difference between two states A and B at equilibrium can be estimated using the cumulative work from state A to state B under nonequilibrium conditions [179, 180]. Jarzynski's equality and its computational implementation are expressed as:

$$\Delta G_{A\to B} \approx W^{x,N}$$

$$\Delta G_{A\to B} = -\frac{1}{\beta}\ln\langle\exp(-\beta W_{A\to B})\rangle_A$$

$$= -\frac{1}{\beta}\ln\sum_{i=1}^{N}\frac{1}{N}\exp\left(-\beta W_{i,A\to B}\right)$$

where $W^{x,N}$ is the exponential average of work for N realizations, and W_i is the nonequilibrium accumulative work (for the i th realization) done onto the system when going from state A to state B in the simulation process. This method generates both statistical and systematic errors [178]. The statistical uncertainty, in terms of standard error, can be calculated using a bootstrap method [181, 182], which is sensitive to the possibility that $W^{x,N}$ might be dominated by one or a few particularly small values of work among all the realizations. In addition, the "fast growth" estimate contains a systematic bias: for finite N, the exponential average of work tends to overestimate $\Delta G_{A\to B}$ by [35]:

$$W^{x,N} - \Delta G_{A\to B} \approx \frac{\beta\sigma_W{}^2}{2N}$$

Normally the statistical errors dominate rather than the systematic errors, and it is suggested that $\sigma_W \geq k_B T$ [178]. 100 cycles of bootstrapping analysis were performed to calculate the statistical error every 200 ps.

The "fast growth" method has been shown to converge to the free energy difference of the two states on small molecular systems both computationally [183] and experimentally [184]. It has been recently incorporated into calculating the product expulsion energy of the exocellulase Cel7A from *T. reesei*, and the result (-14.4 kcal/mol) was qualitatively comparable with that calculated by the free energy perturbation method (-11.2 kcal/mol) [177]. The product expulsion energies calculated using this method were also qualitatively consistent with the experimental results that product inhibition mostly affects exocellulases (such as Cel7A and Cel6A from *T. reesei*) rather than endocellulases (such as Cel7B from *T. reesei* and Cel6B from *Humicola insolens*) [185].

The product inhibition in family 48 cellulases might be related to the composition of amino acid residues at the tunnel exit, particularly the ones that are charged and the ones that form hydrogen bonds with the cellobiose. We designed rational single mutants with the mutation sites along the tunnel exit, aiming at reducing the product expulsion energies so as to reduce the product inhibitory level. To identify the mutation sites, we screened the 40 streams of the SMD simulations of each wildtype cellulase to detect the tunnel exit residues that had long-term strong interactions with the cellobiose. The criteria were: the selected residues needed to be within 4.5Å of the cellobiose for longer than 2 ns out of the 22 ns simulation; the interaction energy between each of these residues and the cellobiose presented either strong VDW interaction (maximum magnitude > 5 kcal/mol) or strong electrostatic energy (maxim magnitude > 17 kcal/mol) to the cellobiose, and such strong interaction occurred at least in 20 out of the 40 SMD simulations. Interestingly, many of the identified mutation sites were conserved or partially conserved among the four cellulases (Table 5.3).

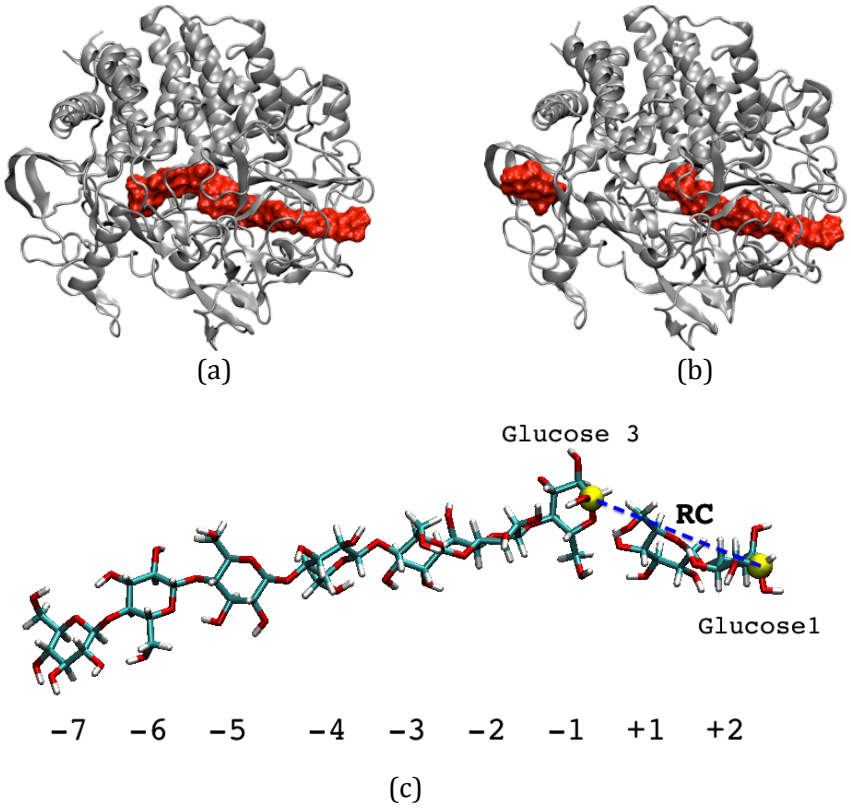


Figure 5.2. Representative structures of the initial (a) and final (b) states of the SMD simulations. (c) The reaction coordinate (RC) was defined to be the distance between the two C1 atoms.

5.3. Results and Discussions

While the structures of the four wildtype family 48 cellulases are similar near the product site except for Cel48, the cellobiose product expulsion free energies vary widely. The calculated product expulsion energies indicate that the product inhibitory level is the highest in CelS, following by Cel48A and CelF, and the lowest in CelA (Figure 5.3). This is qualitatively consistent with the experimental observation that CelA has much higher enzymatic activity than the other three cellulases. Although the product expulsion energy can not explain, alone, the relative activity for these cellulases, it probably makes a large contribution to product inhibition. Compared with the CelA, CelS, and CelF structures, Cel48 exhibits an extra loop structure at the tunnel exit (Figure 5.1), which is a turn composed of the residues from 467 to 471. This part was initially speculated to be a cause for the higher product inhibitory level in Cel48, compared with CelA. However, the Cel48

potential of mean force (PMF) profile showed that the free energy change reached a plateau for reaction coordinate values of ~10 Å, before the cellobiose started to have contact with the turn. Hence, the turn in Cel48 does not seem to affect the level of product inhibition.

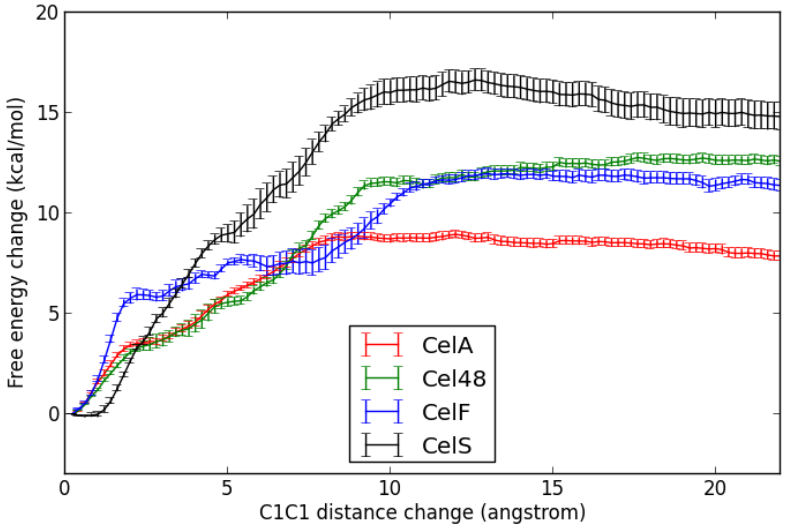


Figure 5.3. The product expulsion energies of the four wildtype family 48 cellulases. The standard errors of all data points were below 1 kcal/mol.

The completion of product escape in the cellulases can be understood from the perspectives of thermodynamics and kinetics. With respect to thermodynamics, the preliminary study on the volume density map of β-D-glucopyranose around the CelF surface illustrated a binding interaction between β-D-glucopyranose and the tunnel exit, near the aromatic residue Trp411 (Chapter 4). There were three events of glucose localization at the tunnel exit pocket over the 100 ns simulation and each localization event lasted for a relatively long time (on the order of ~25ns), suggesting a relatively stable binding of cellulose units to the specific site compared to the protein surface in general. Certainly, a relatively high cellobiose concentration would augment the inhibitory effects on the cellulases. In terms of kinetics, the calculated product expulsion energies were seemingly too high for the cellobiose product to ever escape the tunnel exit. In the native environment, however, the cellobiose is only an intermediate product, and it is subsequently hydrolyzed into glucose by the

cellulase-producing microorganism. For example, in nature, after the solubilization process of crystalline cellulose by the excreted cellulosomes, *C. thermocellum* transported the soluble cellodextrins into the cell plasma via ATP-binding cassette transports [65], and further degraded them into glucoses by cellodextrin and cellobiose phosphorylases [186]. In cellulosic biofuel production, β-glucosidase is commonly thought of as a good candidate for post-cellulase hydrolysis of cellulose. Several studies have shown that the addition of β-glucosidase enhanced the rate of solubilization for crystalline cellulose [187, 188]. In particular, Gefen and others have cleverly integrated a cohesion-fused β-glucosidase into the *C. thermocellum* cellulosome, leading to an increase of cellulose-degradation efficacy by ~2 fold, higher than using a mixture of cellulosome and free β-glucosidase [189]. These studies indirectly suggested that the cellobiose product is able to overcome the product expulsion energy well to be further utilized.

In the SMD simulations, there was no control over the orientation of the celloheptaose and the cellobiose. The celloheptaose maintained its position in the well-defined tunnel, and thus had no significant changes in either the Φ,Ψ-torsional angles of the glycosidic linkages nor the Euler angles. For the cellobiose, analysis of the Φ,Ψ-torsional angles over the course of product escape showed that they were within the same range as that of the relaxed cellobiose in aqueous solution [190]. The Euler angles of the cellobiose varied during this process, since cellobiose is not a rigid body, and such variations were considered averaged out through the multiple SMD simulations.

It is noteworthy that this study might only represent part of the mechanism for product inhibition. For example, the course of product escape was designed to be such that when moving the cellobiose product out of the tunnel exit, the substrate maintained its position in the substrate side of the tunnel, while in the real scenario, both events might happen simultaneously. In that case, the substrate gradually binds to the subsites +1 and +2, reducing the free energy difference between these initial and final states. Nonetheless, previous efforts in guiding the substrate (celloheptamer in this system) to proceed over the active site in the simulations were rather

unsuccessful due to the narrow path at the active site. Also, the crystal structures of the cellulases might not represent all the conformations over the course of product escape. Overall, as the goal was to find mutation sites along the tunnel exit of the cellulases, particularly the sites beyond the subsites +1, to reduce binding to the cellobiose product over its escape, it was decided not to involve the process of substrate proceeding in the simulations.

The composition of amino acid residues at the tunnel exit affects the product expulsion energy. Rational single mutants with the mutation sites along the tunnel exit were designed, aiming at reducing the product expulsion energies. The product expulsion energies of wildtype cellulases and their rational mutants are shown in Table 5.2 and Figure 5.4, and are further presented in a more intuitive way in Table 5.3. The mutants with reduced product expulsion energies indicated possibly good candidates with reduced levels of product inhibition.

Table 5.2. Product expulsion energies of the four family 48 cellulases and their mutants

Cellulase and its Mutation site	Mean of the product expulsion energy (kcal/mol)	Cellulase and its Mutation site	Mean of the product expulsion energy (kcal/mol)
CelA	8.191	CelF	11.478
CelA_R548A	**6.848**	**CelF_R544A**	**8.531**
CelA_D489A	**7.351**	**CelF_R549A**	**7.129**
CelA_D493A	**1.827**	**CelF_D494A**	**8.056**
CelA_D550A	**3.545**	**CelF_D546A**	**9.336**
CelA_E44A	**4.903**	**CelF_E44A**	**4.309**
CelA_E546A	**1.163**	CelF_E542A	15.987
CelA_K547A	13.948		
Cel48	12.635	CelS	15.033
Cel48_R592A	**10.907**	**CelS_R643A**	**10.629**
Cel48_R682A	14.595	**CelS_D490A**	**11.511**
Cel48_D530A	14.335	**CelS_D520A**	**8.968**
Cel48_D534A	20.332	**CelS_E76A**	**4.707**
Cel48_D594A	17.115	CelS_E576A	19.624
Cel48_E38A	19.186		
Cel48_E590A	20.138		
Cel48_E591A	15.126		
Cel48_E593A	13.214		

Note: The mutants in bold have reduced level of product expulsion energy compared to the corresponding wildtype cellulase.

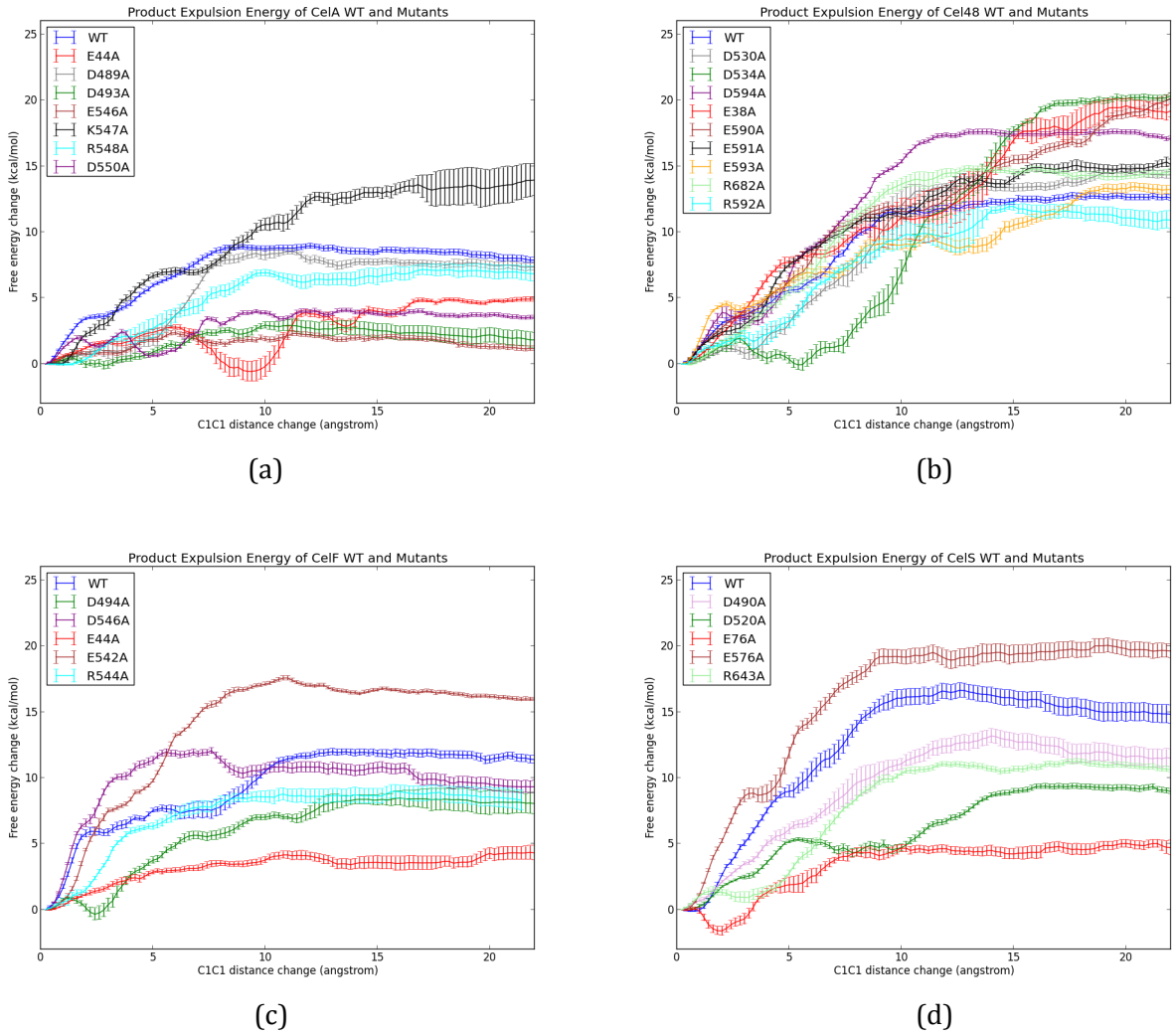


Figure 5.4. Calculated cellobiose product expulsion energy in the wildtype (WT) and rational mutants of CelF (a), Cel48 (b), CelF (c), and CelS (d). Standard error of all data points was below 1 kcal/mol. The homologous mutants of the four cellulases (Table 5.1) are plotted in the same color.

Mutating the tunnel exit residues of the cellulases can change the topological features of the region and thus affect its affinity to the cellobiose product over the course of product escape. In particular, at the initial stage of product escape, the cellobiose is stuck in a narrow pocket (Figure 5.5). On the bottom side of the tunnel exit, the conserved residues including two Arg's, two Glu's, and one Asp in the groups 1, 7, 8, 10, and 11, form a flat vdW surface via multiple salt bridges. These charged residues can form multiple hydrogen bonds with the cellobiose. On the other side, the

conserved residues Trp and Asp in the groups 2 and 5 form stacking interaction and hydrogen bonds with the cellobiose, respectively. It is likely that the flat bottom surface and the Trp residue on the top side function together in stabilizing the substrate prior to its hydrolysis. Substituting the group 8 Arg into Ala reduces the product expulsion energy, the cause for this is likely to be that the Arg associates with three acidic amino acids in the groups 1, 7, and 10, and plays a key role in forming the flat surface. Removal of this Arg truncates the flat surface region and therefore eases the product release. The interaction energy between each of the mutation sites and the cellobiose over the course of product escape is dominated by electrostatic interactions, rather than vdW interactions. Analysis of the intermolecular electrostatic interactions indicates that the acidic residues induce attractive interactions whereas the basic residues of the groups 8 and 11 tend to induce repulsive interactions to the cellobiose (Figure 5.6). The homologous mutation sites, however, did not always follow the same trend in the changes of the product expulsion energies compared to the wildtype enzymes. For example, mutating the group 1 Glu or the group 5 Asp reduces the product expulsion energies in CelA, CelF, and CelS, but increase that in Cel48. Also, mutating the group 7 Glu causes the increase in the product expulsion energies in Cel48, CelF, and CelS, but not CelA.

At the later stage of product escape, the partially conserved group 4 Glu and the charged group 6 residues seem to have impact on the product escape. Interestingly, mutating the group 6 Lys in CelA severely increases the product expulsion energy, and mutating the group 6 Glu in Cel48 also increases the expulsion energy. This combined with the result that each of these residues form attractive interactions with the cellobiose product (Figure 5.6), indicates that the charged residues might assist in passing along the product more easily than other amino acid species. This suggests that including a charged residue at this location, particularly Lys, might facilitate the product escape.

Table 5.3. Product expulsion energies for the four family 48 cellulases and their rational mutants

Group ID	CelA		Cel48		CelF		CelS	
1*	**Glu44**	↘↘	**Glu38**	↗↗	**Glu44**	↘↘	**Glu76**	↘↘
2#	TRP412		TRP411		TRP411		TRP439	
3	Thr463		Gly500		Thr462		**Asp490**	↘
4#	**Asp489**	↘≈	**Asp530**	↗	Asn490		Ser516	
5#	**Asp493**	↘↘	**Asp534**	↗↗	**Asp494**	↘↘	**Asp520**	↘↘
6	**Lys547**	↗↗	**Glu591**	↗	Gln543		Ala577	
7*	**Glu546**	↘↘	**Glu590**	↗↗	**Glu542**	↗↗	**Glu576**	↗↗
8*	**Arg548**	↘	**Arg592**	↘	**Arg544**	↘	Arg578	
9	Ala549		**Glu593**	↗≈	Gly545		Ala579	
10*	**Asp550**	↘↘	**Asp594**	↗↗	**Asp546**	↘	Asp580	
11*	Arg613		**Arg682**	↗	Arg609		**Arg643**	↘

The residues in bold are the ones that possess high interaction energy with the cellobiose product over the course of its escape. ↗ (or ↘) refers to the increase (or decrease) in ΔG compared to the wildtype with the magnitude from 1~3 kcal/mol; ≈ means the changes in ΔG is moderate with the magnitude <1 kcal/mol; ↗↗(or ↘↘) refers to much larger level of ΔG changes with the magnitude > 3kcal/mol. * refers to the residues that form a flat surface on the lower side of the inner tunnel exit. # refers to the residues on the upper side of the tunnel exit (Figure 5.5).

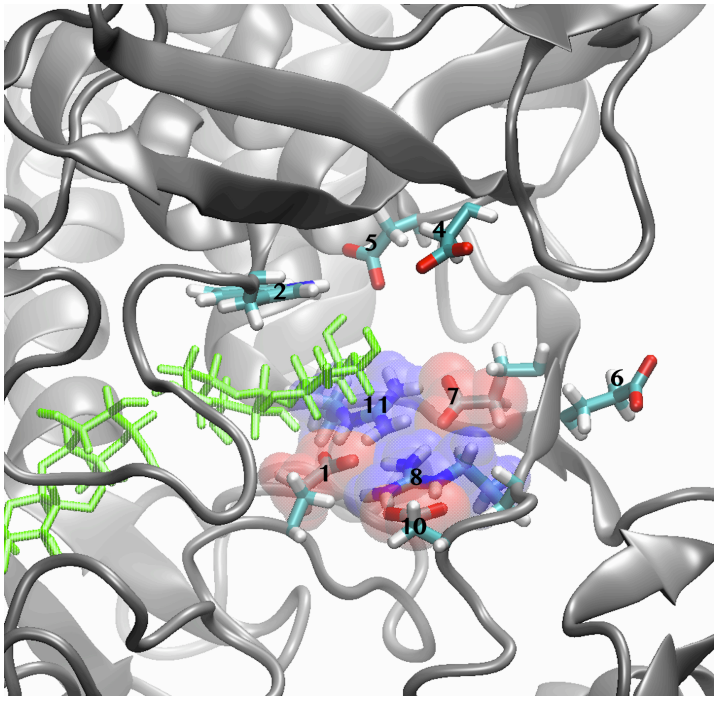


Figure 5.5. The residues that affect the product escape at the tunnel exit of family 48 cellulases. The residues are labeled by their Group IDs. The residue groups 1, 7, 8, 10, and 11 that form a flat surface at the inner part of the tunnel exit are presented in licorice and transparent vdW spheres. The red vdW spheres refer to acidic amino acid residues, and the blue ones refer to basic residues. The representation corresponds to the crystal structure of Cel48. The groups 3 and 9 are not shown since the group 3 is further away from the active site and group 9 is neither conserved nor has a large effect on product escape.

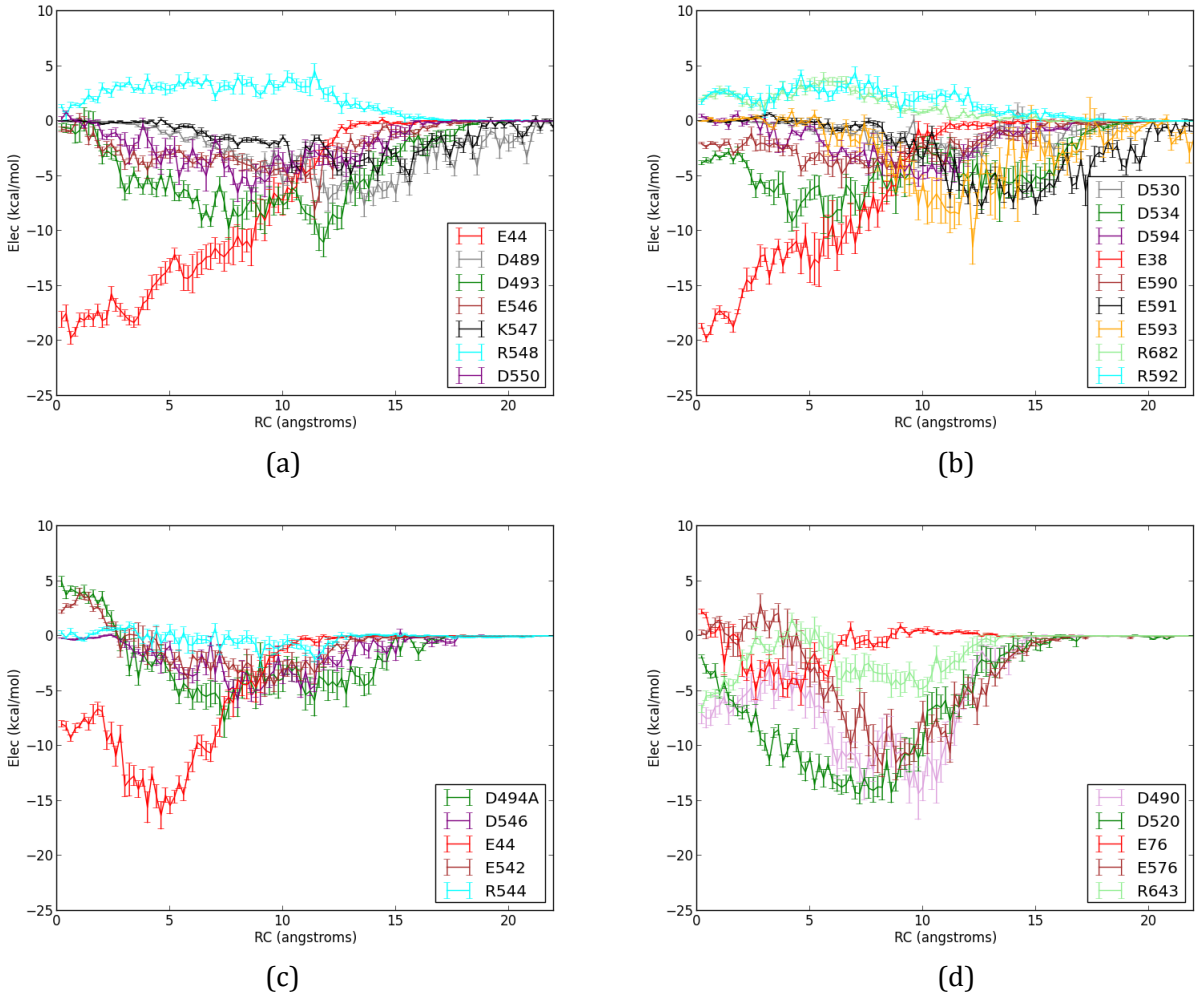


Figure 5.6. The electrostatic interaction between each mutation site and the cellobiose product in CelA (a), Cel48 (b), CelF (c), and CelS (d). The homologous mutants of the four cellulases (Table 5.1) are plotted in the same color. The standard errors are calculated using bootstrapping method.

5.4. Conclusions

Product inhibition in family 48 processive exocellulases is considered one cause of their low catalytic activities. The crystal structures of the cellulases revealed that a cellobiose unit stays in a pocket at the tunnel exit of the cellulases. Also, a preliminary study on the glucose distribution around the surface of a family 48 cellulase, CelF, suggested a higher binding affinity between the cellobiose product and the tunnel exit compared to the protein surface in general. It is hypothesized that the product binds to the tunnel exit of the cellulases, inhibiting its functioning. In this study, the

binding affinity of cellobiose at the tunnel exit of the cellulases was evaluated by free energy calculations of the product expulsion. The calculated product expulsion energies of the four wildtype cellulases seem to be consistent of the experimental results, where CelA has the highest turnover number and Cel48 has the lowest turnover number under comparable conditions. We further designed and evaluated rational single mutants with the mutation sites along the tunnel exit, aiming at reducing the product expulsion energies. Certain single mutants at the conserved or partially conserved residue sites seem to have lower binding affinity to cellobiose compared to the wildtype enzymes. In particular, mutating the residues in the groups 1, 5, and 8 into Ala in most family 48 cellulases might be effective in reducing the product inhibitory effect. Mutating the residues in group 6 into Lys or other charged amino acids might also ease the product escape. Thus, we have identified theoretically plausible mutants of the family 48 cellulases. Further experimental studies are needed to verify the effectiveness of these mutants in improving the turnover number of the cellulases.

CHAPTER 6

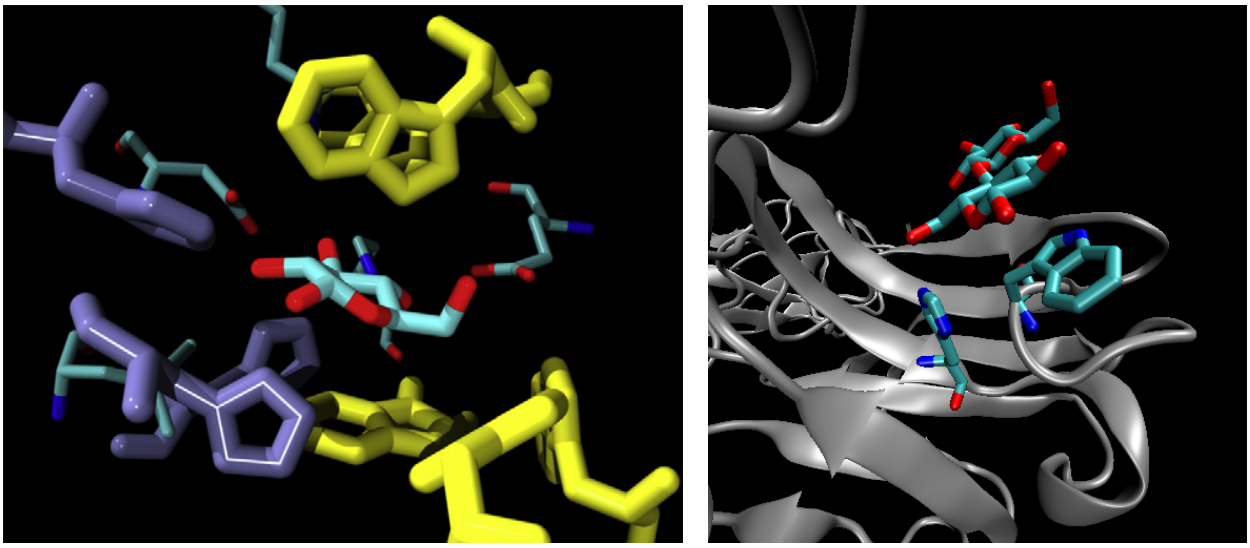
MOLECULAR SIMULATIONS OF INTERACTION OF β-D-GLUCOPYRANOSE WITH IMIDAZOLE IN AQUEOUS SOLUTIONS

6.1. Introduction

Glucose is the structural unit of cellulose. Like many other sugars, it generally behaves as an osmolyte because it contains multiple hydroxyl groups that favor interactions with water molecules through hydrogen bonding. Therefore, in general, glucose would be preferentially excluded from the surfaces of proteins. However, there are a number of proteins that can specifically bind to sugars. For example, many proteins include the glycoside hydrolases such as lysozyme, amylases, and cellulases, the enzymes that utilize carbohydrates as substrates. In particular, most of the processive exocellulases not only contain a catalytic center designed to act on a carbohydrate substrate, but also possess a distinct binding domain designed to anchor the enzyme to its cellulose substrate via strong and specific binding [191]. The various cellulases are particularly interesting examples of sugar binding proteins because of their importance in biomass conversion efforts [192].

In these specific proteins, the sugar-binding sites have been naturally designed to overcome the preference of sugars to fully hydrate and to avoid the surfaces of proteins. Since a number of crystal structures of proteins from each of these classes are available, it is possible from surveying these structures to make some general statements about the types of interactions that favor sugar binding. In particular, the glucose-binding sites commonly contain the amino acids with planar sidechains such as tryptophan, phenylalanine, tyrosine, and histidine. The flat faces of these sidechains are non-hydrogen bonding and thus can be effectively hydrophobic. Their role in promoting the affinity for glucose is to stack their hydrophobic surfaces on the hydrophobic 'tops' and 'bottoms' of the beta anomer of glucose. Particularly, the indole group of Trp sidechain

possesses an extended surface, and it seems to be especially effective for this role as Trp occurs frequently in cellulase binding sites [165, 193]. A study on glucose interacting with the indole group of tryptophan found that the principal mode of interaction was by stacking the H1–H3–H5 hydrophobic triad of the glucose molecule against the planer, hydrophobic face of the indole rings [194]. Another study calculated the free energy landscape for the interaction between indole and β-D-glucopyranose in aqueous solution, indicating that the two species favored stacking interactions, corresponding to a binding energy of ~1.2 KJ/mol [195]. In addition, histidine is also often found in such sites, in spite of its considerably smaller planar surface area. Figure 6.1a shows a glucose binding site containing three His residues [196]. Figure 6.1b presents a carbohydrate binding module of a cellulase, showing a His at the carbohydrate binding site [197]. In addition, a conserved His residue is also found in the active site tunnel of family 48 cellulases [144, 165], very close to the scissile bond of the substrate, as is shown in Figure 6.2. It would be interesting to know whether there is any particular affinity for the imidazole group of His for glucose, which represents the cellulose substrate to some degree as it is the repeating monomer of cellulose; and, if so, if that interaction is mediated by hydrogen bonding, hydrophobic face-to-face stacking, or some other type of interaction.



(a) (b)

Figure 6.1. (a) The binding site of a glucose binding protein with a glucose ligand, illustrating stacking of the sugar against a Trp indole group. Residues within 4.5 Å of the ligand are shown, including 5 Trp's (shown in yellow), the imidazole groups of three His's (shown in blue), and two acid residues. (b) The crystal structure of a carbohydrate binding module of *C. thermocellum* cellulosome containing a Trp and a His within 3.5 Å of the crystallized cellobiose.

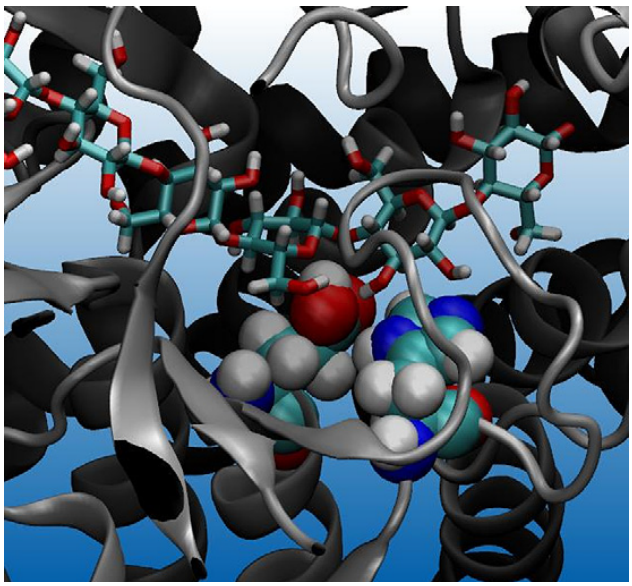


Figure 6.2. The active site of a family 48 cellulase CelF from *Clostridium cellulolyticum*, illustrating the position of a conserved His36 in the active site and the catalytic acid residue Glu55.

Imidazole could potentially interact with glucose both as a planar hydrophobic surface and as a hydrogen-bond acceptor with its N3 nitrogen atom (Figure 6.3). The relative orientations of the

sugar ligand and His imidazole groups in the crystallographic structure of a glucose binding protein (Figure 6.1a) and a carbohydrate binding module (Figure 6.1b) suggest hydrogen-bonding interactions. In CelF, the His36 residue not only is making direct interactions with the substrate, but probably more importantly, may be serving a catalytic role through its proximity to the catalytic acid Glu55, due to its well-known ability to serve as a proton donor and acceptor as imidazolium. In its neutral, imidazole form, the charge distribution, can support strong hydrogen bonding as an acceptor to N3, but probably not to the N1–H group [198].

The present study consisted of a molecular dynamics simulation of imidazole in an aqueous solution of glucose to examine what interactions take place between these two species, as well as between the neutral imidazole molecules themselves. Similar simulations have recently been used to study the interactions of glucose with the Trp residue of melittin from bee venom [194], and with the Tyr residues in a carbohydrate binding module of CBHI from *Trichoderma reesei* [199]. Understanding any interactions that take place between imidazole and glucose could help illuminate what role His sidechains might be playing in the mechanisms of binding and hydrolysis in cellulases like the family 48 cellulases.

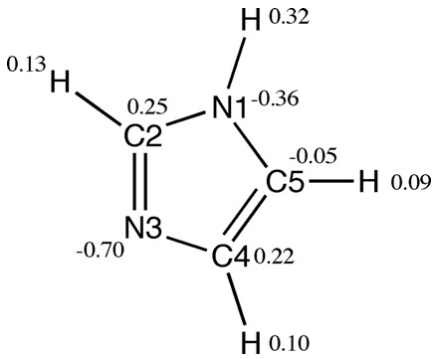


Figure 6.3. The structure of imidazole, the atomic point charges, and the atomic nomenclature used in this study.

6.2. Methods

Molecular dynamics simulation was used to investigate the interaction between glucose and

imidazole in aqueous solution. The primary simulation box contained 30 imidazole molecules and 30 β-D-glucopyranose molecules, with 555 water molecules in a cubic box with the dimension of 29.6 Å in length. This ratio produced concentrations of 3 molal in both the sugar and the imidazole. The initial coordinates for the system were generated by randomly placing each imidazole and glucose molecule into a previously equilibrated box of TIP3P water [200] and removing any water molecules whose oxygen atom was closer than 2.4 Å to any solute heavy atom. Initial configurations were first minimized with 50 steps of steepest descent minimization to remove bad local contacts. Next, the system was heated from 0 to 298 K over a 100 ps period of thermalization. After that, the simulations were run for an additional 15 ns in the canonical NVT ensemble using the CHARMM molecular simulation package [111, 112]. The lengths of the covalent bonds involving hydrogen atoms were kept fixed using the SHAKE algorithm [169]. The Newtonian equations of motions were integrated using a time step of 1 fs. Van der Waals interactions were smoothly truncated on an atom-by-atom basis using switching functions from 10.0 to 12.0 Å. Electrostatic interactions were treated using the particle-mesh Ewald (PME) method [116, 117] with a cutoff value of 12.0 Å, a bspline order of 6, a KAPPA of 0.32, and a grid number of 30 along each dimension. The size of the box was adjusted to 29.6 Å to yield the density of water at 25 °C. Atomic density analyses of the trajectory were displayed using the Visual Molecular Dynamics (VMD) graphics program [201]

In an actual solution of glucose in water, the sugar would quickly undergo tautomerization, giving a 64:36% (β:α) ratio of the pyranoid anomers [202]. In the present study, only the beta anomer was used because the primary interaction of interest is that of the glucose monomers of cellulose, where the glucose monomers are connected via beta glycosidic linkages. Also, only the unprotonated imidazole form was present, representing a solution at neutral pH.

6.3. Results and Discussion

Although imidazole has very high solubility in aqueous solution [203], a weak tendency is observed for imidazole to self-associate at this concentration in the simulation. This is consistent

with the results of a very recent study of imidazole in solution with both the AMBER point charge model and a multipolar expansion model for the electrostatics [204]. Figure 6.4 displays the contoured volume density of imidazole molecules relative to a coordinate frame fixed with respect to a central imidazole, and averaged over all imidazole molecules of all the frames of the simulation. There is a clear tendency for the molecules to pair in solution. Most of these interactions (~70%) are through perpendicular, or T-type, interactions, as seen in Figure 6.5a. A much smaller fraction of interactions, about 7.5%, take place by stacking against one another's planer faces (Fig. 6b), in a manner similar to that seen for caffeine, which is known from osmotic experiments to form such aggregates [24]. Unlike the caffeine case, however, this tendency is much weaker for imidazole and no larger aggregates are seen. A third type of imidazole pairing also has the two molecules perpendicular to one another, but with the N1–H bond of one molecule hydrogen bonded to the N3 atom of the second (Fig. 6c), forming a chain-like type of interaction, accounting for approximately 22.5% of the dimer pairs. Only one imidazole concentration was studied in the present studies, and it has been shown that the imidazole concentration affects the distribution of the imidazole self-interacting relative geometries [204]. Also, system size could affect aggregation in some systems [205], though the previous simulations of similar systems which studied aggregation as a function of system size had no such effects [206].

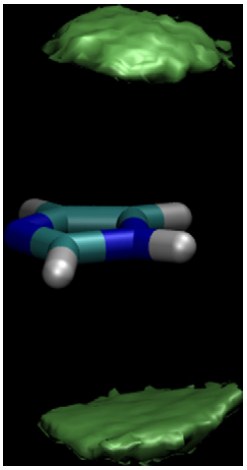


Figure 6.4. Isocontour density surfaces enclosing those regions of space, relative to a frame fixed with respect to the imidazole solute, where the density of other imidazole ring atoms exceeds 3 times the bulk value.

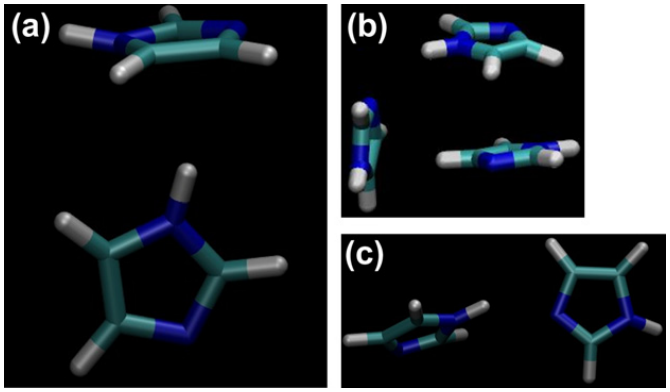


Figure 6.5. The relative geometries for typical imidazole interactions. (a) a T-type interaction; (b) a stacking interaction mediated by a second T-type association; (c) a chain-type interaction.

Liem and coworkers also found self-associations of imidazole species in aqueous solution by MD simulations using both an AMBER classical point-charge model and a classic atomic model with a high-rank multipolar potential [204], the former of which is similar to the CHARMM model used here. They only reported two types of dimers, the chain-like arrangement of Figure 6.5c and the face-to-face stacked arrangement of two of the molecules in Figure 6.5b. They did not report any of the T-type, as shown in Figure 6.5a. However, examination of their density distributions, shown for individual atoms types, reveals that significant pairing of this type occurred in their simulations as well [204].

A weak binding affinity between glucose and imidazole was observed in the simulations. Figure 6.6 displays contours of high density for the ring heavy atoms of glucose, and the two faces of glucose, which are composed of the glucose H1, H3, and H5 protons and the glucose H2 and H4 protons, respectively. As can be seen, there is a high probability of a glucose molecule stacking onto the flat faces of the imidazole solute, in a manner very similar to that previously seen for indole [194, 195], tyrosine [195, 199], and caffeine [207]. This result might not be completely unexpected given these previous studies. However, it is not obvious that glucose would stack in this manner since the imidazole group has a much smaller flat surface area. A strong preference for the β-D-glucopyranose molecules to be orientated with the H1– H3–H5 hydrophobic triad in van der Waals contact with the imidazole ring was also observed, and with the H2 and H4 protons pointing away, and with a very much lower probability for it to be oriented the other way around (Figure 6.6).

Using the same method that was described previously (Chapter 4), the free energy of binding was estimated from the trajectory data by calculating a host–guest type equilibrium constant from the concentration of bound glucose. In this case, the bulk density of each atom selection was calculated to be the number of the selected atoms divided by the volume of the cubic system. The binding energy of glucose with imidazole is shown in Table 6.1. Using the density clouds of the glucose ring heavy atoms, the binding energy of two species are calculated to be

approximately 0.5 kcal/mol, which is lower than that for the face-to-face binding of a Trp indole group with glucose in melittin [5], but is on the same order as computed from glucose–indole binding from potential of mean force calculations [6]. Imidazole–imidazole interactions, computed in the same fashion, are stronger, approximately 0.8 kcal/mol (Table 6.2).

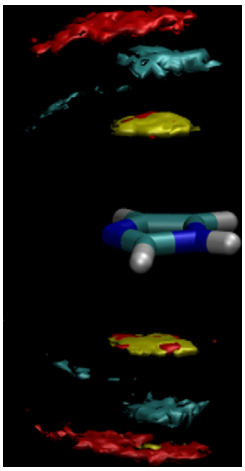


Figure 6.6. Isocontour density surfaces enclosing those regions of space, relative to a frame fixed with respect to the imidazole solute, where the density for the β-D-glucopyranose atoms exceeds the bulk value by a factor of 2.5. Red: the density of the aliphatic protons H2 and H4; yellow: the density of the H1, H3, and H5 aliphatic protons; metallic blue: the density of the ring carbon atoms.

Table 6.1. The binding energy for β-D-glucopyranose pairing with imidazole calculated from the density data, as a function of the contour level selected to define the binding site

Contour level (× bulk density)	K_{eq}	Calculated binding energy (kcal/mol)
2.2	2.38	-0.51
2.4	2.52	-0.55
2.6	2.68	-0.58
2.8	2.82	-0.61

Table 6.2. The binding energy for imidazole–imidazole pairing calculated from the density data, as a function of the contour level selected to define the binding site

Contour level (× bulk density)	K_{eq}	Calculated binding energy (kcal/mol)
3.2	3.50	-0.74
3.4	3.64	-0.76
3.6	3.80	-0.79
3.8	3.95	-0.81

In addition to this hydrophobic stacking, β-D-glucopyranose also interacts with solvated imidazole through hydrogen bonding. Figure 6.7 displays the density of glucose oxygen atoms around the imidazole ring. As can be seen, the density of the glucose ring oxygen atom O5 is only above and below the imidazole faces, resulting from stacking. The hydroxyl oxygen atoms, however, are arrayed around the ring in orientations that indicate hydrogen bonding. The density of these hydroxyl oxygen atoms forms a banana- shaped band arcing around the N3 position in particular, due to its strongly negative partial atomic charge, allowing it to make particularly strong interactions as a hydrogen bond acceptor. The presence of diffuse O6 oxygen atom density around the C–H positions of C2 and C4, but not C5, is harder to explain, but probably is the incidental result of the topological constraint of the glucose molecule—if one of the O1–O4 hydroxyl groups is making a hydrogen bond to N3, the exocyclic hydroxyl group is more likely to be nearby due to topological constraints; this would explain why there is no such diffuse density around C5.

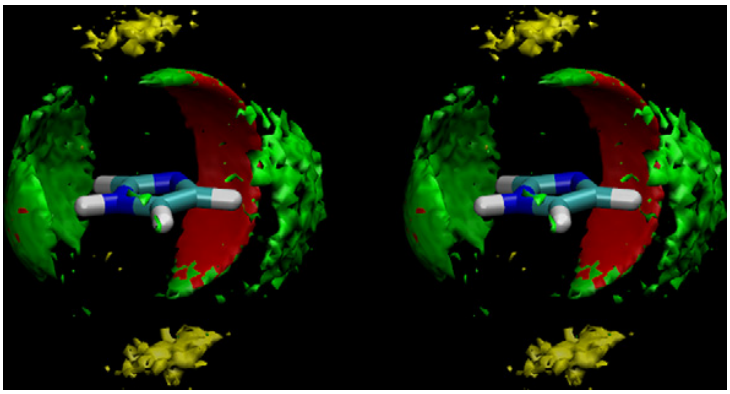
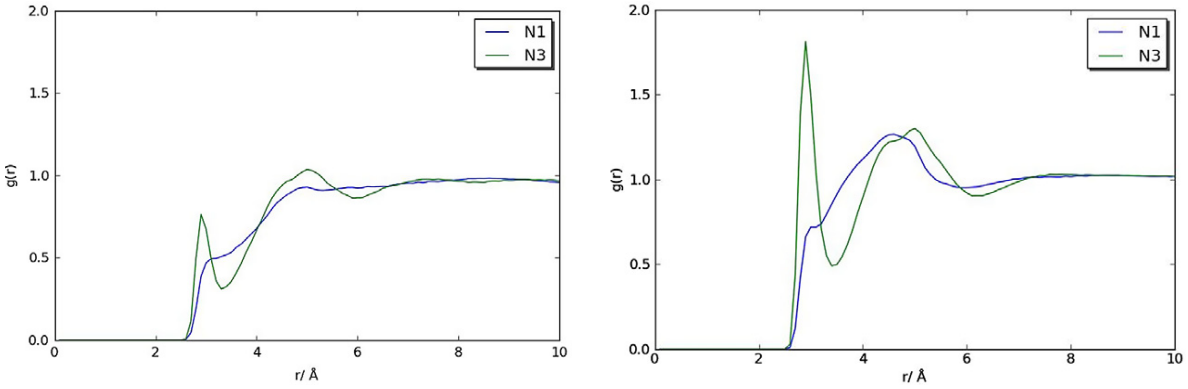


Figure 6.7. The β-D-glucopyranose density around imidazole. Red: the density of O1, O2, O3, and O4 atoms contoured at 3× bulk density; yellow: the density of O5 atoms contoured at 3× bulk density; green: the density of O6 atoms contoured at 3× bulk density. Note particularly the density of the O5 ring atom, which indicates the ring stacking above and below the imidazole plane.

The radial distribution function $g_{NO}(r)$ for sugar hydroxyl oxygen atoms was calculated for both imidazole nitrogen atoms, as is shown in Figure 6.8a. The distribution around the N3 nitrogen atom shows a weak hydrogen-bond interaction, while there is no such peak in the $g_{NO}(r)$ for the N1

position. These distributions can also be compared to similar radial distribution functions for water oxygen atoms around these nitrogen atoms, as shown in Figure 6.8b. These plots show that the sugar hydroxyl oxygen atoms are hydrogen bonding to these two atoms in the same manner as do the water molecules. Table 6.3 lists the number of hydrogen bonds to water and to the sugar oxygen atoms calculated from the trajectory using as a definition a distance cutoff of 3.4 Å and an angle cutoff of 150° or greater. The unprotonated N3 atom makes, on average, approximately 1.4 hydrogen bonds to water, and approximately 0.2 to a sugar hydroxyl group, so that this atom is generally making a little less than two hydrogen bonds. The near absence of hydrogen bonding for the N–H group of N1 is due to the very low charges of these atoms, since the polarity of this group is insufficient to compete for partners with the much stronger water– water hydrogen bonds.



(a) (b)

Figure 6.8. The radial distribution functions of glucose oxygen atoms around the two nitrogen atoms of imidazole (a), and that of water oxygen atoms around the two nitrogen atoms of imidazole.

Table 6.3. The number of hydrogen bonds made by the imidazole nitrogen atoms to both water and glucose hydroxyl groups

Atom	Hydrogen bonds to glucose	Hydrogen bounds to water
N1	0.022	0.144
N3	0.214	1.393

6.4. Conclusions

The present simulations find that in aqueous solution there is a significant tendency for glucose to associate with imidazole both through face-to-face stacking and through hydrogen bonding between the glucose hydroxyl groups and the imidazole N3 atom. Both types of interactions could potentially play a role in the affinity of glucose for protein binding sites, although the binding interaction is quite weak, less than kT at room temperature, and by itself would not suffice to position an osmolyte like glucose into a protein site. Somewhat surprisingly, an even greater tendency for imidazole to self-associate was observed, by face-to-face stacking, chain-type interactions, and much more commonly by T-type interactions. These latter two associations are presumably due to favorable electrostatic interactions, whereas stacking interactions both between imidazole and imidazole, and between imidazole and glucose, are probably driven by hydrophobic pairing due to the weak hydration of these surfaces. Understanding such interactions could prove useful in designing improvements in binding site energetics in practical systems like cellulases, where better enzymes are needed for biomass conversion applications.

REFERNCES

1. Lange, J.P., *Lignocellulose conversion: an introduction to chemistry, process and economics.* Biofuels Bioproducts & Biorefining-Biofpr, 2007. **1**(1): p. 39-48.
2. Zhang, Y.H. and L.R. Lynd, *Toward an aggregated understanding of enzymatic hydrolysis of cellulose: noncomplexed cellulase systems.* Biotechnol Bioeng, 2004. **88**(7): p. 797-824.
3. Atalla, R.H. and D.L. Vanderhart, *Native cellulose: a composite of two distinct crystalline forms.* Science, 1984. **223**(4633): p. 283-5.
4. O'Sullivan, A.C., *Cellulose: the structure slowly unravels.* Cellulose, 1997. **4**: p. 173- 207.
5. Boerjan, W., J. Ralph, and M. Baucher, *Lignin biosynthesis.* Annu Rev Plant Biol, 2003. **54**: p. 519-46.
6. Ralph J, L.K., Brunow G, Lu F, Kim H, Schatz PF, Marita JM, Hatfield RD, Ralph SA, Christensen JH, Boerjan W., *Lignins: natural polymers from oxidative coupling of 4-hydrxophenylpropanoids.* Phytochem Rev., 2004. **3**: p. 29-60.
7. Scheller, H.V. and P. Ulvskov, *Hemicelluloses.* Annual Review of Plant Biology, Vol 61, 2010. **61**: p. 263-289.
8. Rubin, E.M., *Genomics of cellulosic biofuels.* Nature, 2008. **454**(7206): p. 841-5.
9. Segal, L., et al., *An empirical method for estimating the degree of crystallinity of native cellulose using the X-ray diffractometer.* Textile Research Journal, 1959. **29**(10): p. 786-794.
10. Lynd, L.R., et al., *Microbial cellulose utilization: fundamentals and biotechnology.* Microbiol Mol Biol Rev, 2002. **66**(3): p. 506-77, table of contents.
11. Bergenstrahle, M., et al., *Simulation studies of the insolubility of cellulose.* Carbohydrate Research, 2010. **345**(14): p. 2060-2066.
12. Ioelovich, M., *Study of Cellulose Interaction with Concentrated Solutions of Sulfuric Acid.* ISRN Chemical Engineering, 2012. **101**: p. Article ID 428974.
13. Jeong, T.-S.O., K-K ;, *Behaviors of Glucose Decomposition during Dilute-Acid Hydrolysis of Lignocellulosic Biomass.* Korean Society for Biotechnology and Bioengineering, 2009. **24**(3): p. 267-272.
14. Hong, Y., et al., *Impact of cellulase production on environmental and financial metrics for lignocellulosic ethanol.* Biofuels Bioproducts & Biorefining-Biofpr, 2013. **7**(3): p. 303-313.
15. Weiss, N., et al., *Enzymatic lignocellulose hydrolysis: Improved cellulase productivity by insoluble solids recycling.* Biotechnology for Biofuels, 2013. **6**.
16. Hisano, H., R. Nandakumar, and Z.Y. Wang, *Genetic modification of lignin biosynthesis for improved biofuel production.* In Vitro Cellular & Developmental Biology-Plant, 2009. **45**(3): p. 306-313.
17. Sun, Y. and J. Cheng, *Hydrolysis of lignocellulosic materials for ethanol production: a review.* Bioresource technology, 2002. **83**(1): p. 1-11.
18. Hendriks, A.T.W.M. and G. Zeeman, *Pretreatments to enhance the digestibility of lignocellulosic biomass.* Bioresource Technology, 2009. **100**(1): p. 10-18.
19. Zhang, Y.H.P., et al., *Fractionating recalcitrant lignocellulose at modest reaction conditions.* Biotechnology and Bioengineering, 2007. **97**(2): p. 214-223.
20. Chundawat, S.P.S., et al., *Restructuring the Crystalline Cellulose Hydrogen Bond Network Enhances Its Depolymerization Rate.* Journal of the American Chemical Society, 2011. **133**(29): p. 11163-11174.
21. Bellesia, G., et al., *Probing the Early Events Associated with Liquid Ammonia Pretreatment of Native Crystalline Cellulose.* Journal of Physical Chemistry B, 2011. **115**(32): p. 9782-9788.
22. Wang, H., G. Gurau, and R.D. Rogers, *Ionic liquid processing of cellulose.* Chemical Society Reviews, 2012. **41**(4): p. 1519-1537.

23. Lee, S.H., et al., *Ionic Liquid-Mediated Selective Extraction of Lignin From Wood Leading to Enhanced Enzymatic Cellulose Hydrolysis.* Biotechnology and Bioengineering, 2009. **102**(5): p. 1368-1376.
24. Li, C.L., et al., *Comparison of dilute acid and ionic liquid pretreatment of switchgrass: Biomass recalcitrance, delignification and enzymatic saccharification.* Bioresource Technology, 2010. **101**(13): p. 4900-4906.
25. Silverstein, R.A., et al., *A comparison of chemical pretreatment methods for improving saccharification of cotton stalks.* Bioresource Technology, 2007. **98**(16): p. 3000-3011.
26. Wang, Z., et al., *Sodium hydroxide pretreatment and enzymatic hydrolysis of coastal Bermuda grass.* Bioresource technology, 2010. **101**(10): p. 3583-3585.
27. Mikkola, J.P., et al., *Ultrasound enhancement of cellulose processing in ionic liquids: from dissolution towards functionalization.* Green Chemistry, 2007. **9**(11): p. 1229-1237.
28. Sinnott, M.L., *Catalytic Mechanisms of Enzymatic Glycosyl Transfer.* Chemical Reviews, 1990. **90**(7): p. 1171-1202.
29. Lynd, L.R., et al., *Microbial cellulose utilization: Fundamentals and biotechnology.* Microbiology and Molecular Biology Reviews, 2002. **66**(3): p. 506-+.
30. Henrissat, B., *A Classification of Glycosyl Hydrolases Based on Amino-Acid-Sequence Similarities.* Biochemical Journal, 1991. **280**: p. 309-316.
31. Henrissat, B. and A. Bairoch, *New Families in the Classification of Glycosyl Hydrolases Based on Amino-Acid-Sequence Similarities.* Biochemical Journal, 1993. **293**: p. 781-788.
32. Henrissat, B. and A. Bairoch, *Updating the sequence-based classification of glycosyl hydrolases.* Biochemical Journal, 1996. **316**: p. 695-696.
33. Chothia, C. and A.M. Lesk, *The relation between the divergence of sequence and structure in proteins.* EMBO J, 1986. **5**(4): p. 823-6.
34. Gebler, J., et al., *Stereoselective hydrolysis catalyzed by related beta-1,4-glucanases and beta-1,4-xylanases.* J Biol Chem, 1992. **267**(18): p. 12559-61.
35. Vinzant, T.B., et al., *Fingerprinting Trichoderma reesei hydrolases in a commerical cellulase preparation.* Applied Biochemistry and Biotechnology, 2001. **91-3**: p. 99-107.
36. Kubicek, C.P. and M.E. Penttila, *Regulation of production of plant polysaccharide degrading enzymes by Trichoderma*, in *Trichoderma and gliocladium*, C.P. Kubicek and G.E. Harman, Editors. 1998, Taylor & Francis: London ; Bristol, PA.
37. Divne, C., et al., *Crystallization and preliminary X-ray studies on the core proteins of cellobiohydrolase I and endoglucanase I from Trichoderma reesei.* J Mol Biol, 1993. **234**(3): p. 905-7.
38. Davies, G.J., et al., *Oligosaccharide specificity of a family 7 endoglucanase: insertion of potential sugar-binding subsites.* J Biotechnol, 1997. **57**(1-3): p. 91-100.
39. Divne, C., et al., *The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from Trichoderma reesei.* Science, 1994. **265**(5171): p. 524-8.
40. Sandgren, M., et al., *The X-ray crystal structure of the Trichoderma reesei family 12 endoglucanase 3, Cel12A, at 1.9 angstrom resolution.* Journal of Molecular Biology, 2001. **308**(2): p. 295-310.
41. Schulein, M., *Enzymatic properties of cellulases from Humicola insolens.* Journal of Biotechnology, 1997. **57**(1-3): p. 71-81.
42. Boisset, C., et al., *Optimized mixtures of recombinant Humicola insolens cellulases for the biodegradation of crystalline cellulose.* Biotechnology and Bioengineering, 2001. **72**(3): p. 339-345.
43. Broda, P., et al., *Phanerochaete-Chrysosporium and Its Natural Substrate.* Fems Microbiology Reviews, 1994. **13**(2-3): p. 189-196.

44. Covert, S.F., A. Vandenwymelenberg, and D. Cullen, *Structure, Organization, and Transcription of a Cellobiohydrolase Gene-Cluster from Phanerochaete-Chrysosporium.* Applied and Environmental Microbiology, 1992. **58**(7): p. 2168-2175.
45. Chaudhary, P., N.N. Kumar, and D.N. Deobagkar, *The glucanases of Cellulomonas.* Biotechnol Adv, 1997. **15**(2): p. 315-31.
46. Irwin, D., et al., *Roles of the catalytic domain and two cellulose binding domains of Thermomonospora fusca E4 in cellulose hydrolysis.* Journal of Bacteriology, 1998. **180**(7): p. 1709-1714.
47. Irwin, D.C., et al., *Activity studies of eight purified cellulases: Specificity, synergism, and binding domain effects.* Biotechnol Bioeng, 1993. **42**(8): p. 1002-13.
48. Ali, B.R., et al., *Cellulases and hemicellulases of the anaerobic fungus Piromyces constitute a multiprotein cellulose-binding complex and are encoded by multigene families.* FEMS Microbiol Lett, 1995. **125**(1): p. 15-21.
49. Ljungdahl, L.G., *The cellulase/hemicellulase system of the anaerobic fungus Orpinomyces PC-2 and aspects of its applied use.* Ann N Y Acad Sci, 2008. **1125**: p. 308-21.
50. Bayer, E.A., R. Kenig, and R. Lamed, *Adherence of Clostridium thermocellum to cellulose.* J Bacteriol, 1983. **156**(2): p. 818-27.
51. Lamed, R., E. Setter, and E.A. Bayer, *Characterization of a cellulose-binding, cellulase-containing complex in Clostridium thermocellum.* J Bacteriol, 1983. **156**(2): p. 828-36.
52. Gaudin, C., et al., *CelE, a multidomain cellulase from Clostridium cellulolyticum: a key enzyme in the cellulosome?* J Bacteriol, 2000. **182**(7): p. 1910-5.
53. Sabathe, F., A. Belaich, and P. Soucaille, *Characterization of the cellulolytic complex (cellulosome) of Clostridium acetobutylicum.* FEMS Microbiol Lett, 2002. **217**(1): p. 15-22.
54. Doi, R.H., et al., *Cellulosome and noncellulosomal cellulases of Clostridium cellulovorans.* Extremophiles, 1998. **2**(2): p. 53-60.
55. Ding, S.Y., et al., *Cellulosomal scaffoldin-like proteins from Ruminococcus flavefaciens.* J Bacteriol, 2001. **183**(6): p. 1945-53.
56. Kirby, J., et al., *Dockerin-like sequences in cellulases and xylanases from the rumen cellulolytic bacterium Ruminococcus flavefaciens.* FEMS Microbiol Lett, 1997. **149**(2): p. 213-9.
57. Ohara, H., et al., *Characterization of the cellulolytic complex (cellulosome) from Ruminococcus albus.* Biosci Biotechnol Biochem, 2000. **64**(2): p. 254-60.
58. Millward-Sadler, S.J., et al., *Evidence that the Piromyces gene family encoding endo-1,4-mannanases arose through gene duplication.* FEMS Microbiol Lett, 1996. **141**(2-3): p. 183-8.
59. Zhou, L., et al., *Intronless celB from the anaerobic fungus Neocallimastix patriciarum encodes a modular family A endoglucanase.* Biochem J, 1994. **297 (Pt 2)**: p. 359-64.
60. Fontes, C.M. and H.J. Gilbert, *Cellulosomes: highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates.* Annu Rev Biochem, 2010. **79**: p. 655-81.
61. Doi, R.H. and A. Kosugi, *Cellulosomes: Plant-cell-wall-degrading enzyme complexes.* Nature Reviews Microbiology, 2004. **2**(7): p. 541-551.
62. Shoham, Y., R. Lamed, and E.A. Bayer, *The cellulosome concept as an efficient microbial strategy for the degradation of insoluble polysaccharides.* Trends Microbiol, 1999. **7**(7): p. 275-81.
63. Bayer, E.A., et al., *Cellulose, cellulases and cellulosomes.* Current Opinion in Structural Biology, 1998. **8**(5): p. 548-557.
64. Gerngross, U.T., et al., *Sequencing of a Clostridium-Thermocellum Gene (Cipa) Encoding the Cellulosomal Sl-Protein Reveals an Unusual Degree of Internal Homology (Vol 8, Pg 325, 1993).* Molecular Microbiology, 1993. **10**(5): p. 1155-1155.
65. Nataf, Y., et al., *Cellodextrin and Laminaribiose ABC Transporters in Clostridium thermocellum.* Journal of Bacteriology, 2009. **191**(1): p. 203-209.

66. Leibovitz, E. and P. Beguin, *A new type of cohesin domain that specifically binds the dockerin domain of the Clostridium thermocellum cellulosome-integrating protein CipA.* J Bacteriol, 1996. **178**(11): p. 3077-84.
67. Lemaire, M., et al., *OlpB, a new outer layer protein of Clostridium thermocellum, and binding of its S-layer-like domains to components of the cell envelope.* J Bacteriol, 1995. **177**(9): p. 2451-9.
68. Dror, T.W., et al., *Regulation of expression of scaffoldin-related genes in Clostridium thermocellum.* J Bacteriol, 2003. **185**(17): p. 5109-16.
69. Raman, B., et al., *Impact of pretreated Switchgrass and biomass carbohydrates on Clostridium thermocellum ATCC 27405 cellulosome composition: a quantitative proteomic analysis.* PLoS One, 2009. **4**(4): p. e5271.
70. Ng, T.K. and J.G. Zeikus, *Comparison of Extracellular Cellulase Activities of Clostridium thermocellum LQRI and Trichoderma reesei QM9414.* Appl Environ Microbiol, 1981. **42**(2): p. 231-40.
71. Zverlov, V.V., et al., *Mutations in the scaffoldin gene, cipA, of Clostridium thermocellum with impaired cellulosome formation and cellulose hydrolysis: insertions of a new transposable element, IS1447, and implications for cellulase synergism on crystalline cellulose.* J Bacteriol, 2008. **190**(12): p. 4321-7.
72. Miras, I., et al., *Mapping by site-directed mutagenesis of the region responsible for cohesin-dockerin interaction on the surface of the seventh cohesin domain of Clostridium thermocellum CipA.* Biochemistry, 2002. **41**(7): p. 2115-2119.
73. Carvalho, A.L., et al., *Evidence for a dual binding mode of dockerin modules to cohesins.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(9): p. 3089-3094.
74. Carvalho, A.L., et al., *Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(24): p. 13809-13814.
75. Schaeffer, F., et al., *Duplicated dockerin subdomains of Clostridium thermocellum endoglucanase CelD bind to a cohesin domain of the scaffolding protein CipA with distinct thermodynamic parameters and a negative cooperativity.* Biochemistry, 2002. **41**(7): p. 2106-2114.
76. Adams, J.J., et al., *Mechanism of bacterial cell-surface attachment revealed by the structure of cellulosomal type II cohesin-dockerin complex.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(2): p. 305-310.
77. Wilson, D.B. and M. Kostylev, *Cellulase processivity.* Methods Mol Biol, 2012. **908**: p. 93-9.
78. Kostylev, M. and D.B. Wilson, *A two-parameter kinetic model based on a time-dependent activity coefficient accurately describes enzymatic cellulose digestion.* Biochemistry, 2013. **52**(33): p. 5656–5664.
79. Vuong, T.V. and D.B. Wilson, *Processivity, Synergism, and Substrate Specificity of Thermobifida fusca Cel6B.* Applied and Environmental Microbiology, 2009. **75**(21): p. 6655-6661.
80. Liu, W., et al., *Engineering of Clostridium phytofermentans Endoglucanase Cel5A for improved thermostability.* Appl Environ Microbiol, 2010. **76**(14): p. 4914-7.
81. Irwin, D., et al., *Roles of the catalytic domain and two cellulose binding domains of Thermomonospora fusca E4 in cellulose hydrolysis.* J Bacteriol, 1998. **180**(7): p. 1709-14.
82. Sakon, J., et al., *Structure and mechanism of endo/exocellulase E4 from Thermomonospora fusca.* Nat Struct Biol, 1997. **4**(10): p. 810-8.
83. Ghose, T., *Measurement of cellulase activities.* Pure Appl Chem, 1987. **59**(2): p. 257-268.
84. Decker, S.R., et al., *Automated filter paper assay for determination of cellulase activity.* Appl Biochem Biotechnol, 2003. **105 -108**: p. 689-703.

85. Irwin, D.C., S. Zhang, and D.B. Wilson, *Cloning, expression and characterization of a family 48 exocellulase, Cel48A, from Thermobifida fusca.* Eur J Biochem, 2000. **267**(16): p. 4988-97.
86. Kostylev, M. and D.B. Wilson, *Determination of the Catalytic Base in Family 48 Glycosyl Hydrolases.* Applied and Environmental Microbiology, 2011. **77**(17): p. 6274-6276.
87. Bezerra, R.M. and A.A. Dias, *Discrimination among eight modified michaelis-menten kinetics models of cellulose hydrolysis with a large range of substrate/enzyme ratios: inhibition by cellobiose.* Appl Biochem Biotechnol, 2004. **112**(3): p. 173-84.
88. Bezerra, R.M. and A.A. Dias, *Enzymatic kinetic of cellulose hydrolysis: inhibition by ethanol and cellobiose.* Appl Biochem Biotechnol, 2005. **126**(1): p. 49-59.
89. Gruno, M., et al., *Inhibition of the Trichoderma reesei cellulases by cellobiose is strongly dependent on the nature of the substrate.* Biotechnol Bioeng, 2004. **86**(5): p. 503-11.
90. Kruus, K., et al., *Product inhibition of the recombinant CelS, an exoglucanase component of the Clostridium thermocellum cellulosome.* Appl Microbiol Biotechnol, 1995. **44**(3-4): p. 399-404.
91. Lamed, R., Kenig, R., Setter, E., and Bayer, E. A., *Major characteristics of the cellulolytic system of Clostridium thermocellum coincide with those of the purified cellulosome.* Enzyme Microb. Technol., 1985. **7**: p. 37-41.
92. Morag, E., et al., *Isolation and properties of a major cellobiohydrolase from the cellulosome of Clostridium thermocellum.* J Bacteriol, 1991. **173**(13): p. 4155-62.
93. Bu, L.N., M. R.; Shirts, M. R.; Stahlberg, J.; Himmel, M. E.; Crowley, M. F.; Beckham, G. T., *Product Binding Varies Dramatically between Processive and Nonprocessive Cellulase Enzymes.* Journal of Biological Chemistry, 2012. **287**(29): p. 24807-24813
94. Baker, J.O., et al., *Catalytically enhanced endocellulase Cel5A from Acidothermus cellulolyticus.* Appl Biochem Biotechnol, 2005. **121-124**: p. 129-48.
95. Percival Zhang, Y.H., M.E. Himmel, and J.R. Mielenz, *Outlook for cellulase improvement: screening and selection strategies.* Biotechnol Adv, 2006. **24**(5): p. 452-81.
96. Zhang, S., B.K. Barr, and D.B. Wilson, *Effects of noncatalytic residue mutations on substrate specificity and ligand binding of Thermobifida fusca endocellulase cel6A.* Eur J Biochem, 2000. **267**(1): p. 244-52.
97. Kim, Y.S., H.C. Jung, and J.G. Pan, *Bacterial cell surface display of an enzyme library for selective screening of improved cellulase variants.* Appl Environ Microbiol, 2000. **66**(2): p. 788-93.
98. Wang, T., et al., *Directed evolution for engineering pH profile of endoglucanase III from Trichoderma reesei.* Biomol Eng, 2005. **22**(1-3): p. 89-94.
99. Heinzelman, P., et al., *A family of thermostable fungal cellulases created by structure-guided recombination.* Proc Natl Acad Sci U S A, 2009. **106**(14): p. 5610-5.
100. Levine, S.E., et al., *A mechanistic model for rational design of optimal cellulase mixtures.* Biotechnol Bioeng, 2011. **108**(11): p. 2561-70.
101. Gefen, G., et al., *Enhanced cellulose degradation by targeted integration of a cohesin-fused beta-glucosidase into the Clostridium thermocellum cellulosome.* Proc Natl Acad Sci U S A, 2012. **109**(26): p. 10298-303.
102. Cornell, W.D., et al., *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995).* Journal of the American Chemical Society, 1996. **118**(9): p. 2309-2309.
103. MacKerell, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* Journal of Physical Chemistry B, 1998. **102**(18): p. 3586-3616.
104. Oostenbrink, C., et al., *A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6.* J Comput Chem, 2004. **25**(13): p. 1656-76.

105. Jorgensen, W.L., D.S. Maxwell, and J. TiradoRives, *Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids.* Journal of the American Chemical Society, 1996. **118**(45): p. 11225-11236.
106. Guvench, O. and A.D. MacKerell, Jr., *Comparison of protein force fields for molecular dynamics simulations.* Methods Mol Biol, 2008. **443**: p. 63-88.
107. Mackerell, A.D., M. Feig, and C.L. Brooks, *Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations.* Journal of Computational Chemistry, 2004. **25**(11): p. 1400-1415.
108. Guvench, O., et al., *Additive Empirical Force Field for Hexopyranose Monosaccharides.* Journal of Computational Chemistry, 2008. **29**(15): p. 2543-2564.
109. Vanommeslaeghe, K., et al., *CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields.* Journal of Computational Chemistry, 2010. **31**(4): p. 671-690.
110. Jorgensen, W.L., et al., *Comparison of Simple Potential Functions for Simulating Liquid Water.* Journal of Chemical Physics, 1983. **79**(2): p. 926-935.
111. Brooks, B.R., et al., *CHARMM: The Biomolecular Simulation Program.* Journal of Computational Chemistry, 2009. **30**(10): p. 1545-1614.
112. Brooks, B.R., et al., *Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations.* Journal of Computational Chemistry, 1983. **4**(2): p. 187-217.
113. D.A. Case, T.A.D., T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A.W. Goetz, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, R.M. Wolf, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P.A. Kollman, *AMBER 12.* 2012.
114. Crowley, M.F., M.J. Williamson, and R.C. Walker, *CHAMBER: Comprehensive Support for CHARMM Force Fields Within the AMBER Software.* International Journal of Quantum Chemistry, 2009. **109**(15): p. 3767-3772.
115. Steinbach, P.J. and B.R. Brooks, *New Spherical-Cutoff Methods for Long-Range Forces in Macromolecular Simulation.* Journal of Computational Chemistry, 1994. **15**(7): p. 667-683.
116. Darden, T., D. York, and L. Pedersen, *Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems.* Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.
117. Essmann, U., et al., *A Smooth Particle Mesh Ewald Method.* Journal of Chemical Physics, 1995. **103**(19): p. 8577-8593.
118. Ryckaert, J.-P., G. Ciccotti, and H.J.C. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.* Journal of Computational Physics, 1977. **23**(3): p. 327-341.
119. Hunenberger, P., *Thermostat algorithms for molecular dynamics simulations.* Advanced Computer Simulation Approaches for Soft Matter Sciences I, 2005. **173**: p. 105-147.
120. Berendsen, H.J.C., et al., *Molecular-Dynamics with Coupling to an External Bath.* Journal of Chemical Physics, 1984. **81**(8): p. 3684-3690.
121. Andersen, H.C., *Molecular-Dynamics Simulations at Constant Pressure and-or Temperature.* Journal of Chemical Physics, 1980. **72**(4): p. 2384-2393.
122. Hoover, W.G., *Canonical Dynamics - Equilibrium Phase-Space Distributions.* Physical Review A, 1985. **31**(3): p. 1695-1697.
123. Nose, S., *A Unified Formulation of the Constant Temperature Molecular-Dynamics Methods.* Journal of Chemical Physics, 1984. **81**(1): p. 511-519.
124. Nose, S., *A Molecular-Dynamics Method for Simulations in the Canonical Ensemble.* Molecular Physics, 1984. **52**(2): p. 255-268.

125. Van Gunsteren, W.F. and H.J.C. Berendsen, *A LEAP-FROG ALGORITHM FOR STOCHASTIC DYNAMICS.* Molecular Simulation, 1988. **1**(3): p. 173-185.
126. Schneider, T. and E. Stoll, *MOLECULAR-DYNAMICS STUDY OF 2ND SOUND.* Physical Review B, 1978. **18**(12): p. 6468-6482.
127. Schneider, T. and E. Stoll, *MOLECULAR-DYNAMICS STUDY OF A 3-DIMENSIONAL ONE-COMPONENT MODEL FOR DISTORTIVE PHASE-TRANSITIONS.* Physical Review B, 1978. **17**(3): p. 1302-1322.
128. Hoover, W.G., *Constant-pressure equations of motion.* Physical Review A, 1986. **34**(3): p. 2499-2500
129. Nose, S. and M.L. Klein, *Constant Pressure Molecular-Dynamics for Molecular-Systems.* Molecular Physics, 1983. **50**(5): p. 1055-1076.
130. Parrinello, M. and A. Rahman, *Polymorphic Transitions in Single-Crystals - a New Molecular-Dynamics Method.* Journal of Applied Physics, 1981. **52**(12): p. 7182-7190.
131. Paci, E. and M. Marchi, *Constant-pressure molecular dynamics techniques applied to complex molecular systems and solvated proteins.* Journal of Physical Chemistry, 1996. **100**(10): p. 4314-4322.
132. Feller, S.E., et al., *Constant-Pressure Molecular-Dynamics Simulation - the Langevin Piston Method.* Journal of Chemical Physics, 1995. **103**(11): p. 4613-4621.
133. Uberuaga, B.P., M. Anghel, and A.F. Voter, *Synchronization of trajectories in canonical molecular-dynamics simulations: Observation, explanation, and exploitation.* Journal of Chemical Physics, 2004. **120**(14): p. 6363-6374.
134. Sindhikara, D.J., et al., *Bad Seeds Sprout Perilous Dynamics: Stochastic Thermostat Induced Trajectory Synchronization in Biomolecules.* Journal of Chemical Theory and Computation, 2009. **5**(6): p. 1624-1631.
135. Allen, M.P. and D.J. Tildesley, *Computer simulation of liquids*. 1989: Oxford university press.
136. Souaille, M. and B.t. Roux, *Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations.* Computer physics communications, 2001. **135**(1): p. 40-57.
137. Hendrix, D.A. and C. Jarzynski, *A "fast growth" method of computing free energy differences.* The Journal of Chemical Physics, 2001. **114**: p. 5974.
138. Jarzynski, C., *A nonequilibrium equality for free energy differences.* arXiv preprint cond-mat/9610209, 1996.
139. Mason, P.E., et al., *Glucose interactions with a model peptide.* Proteins, 2011. **79**(7): p. 2224-32.
140. Pratt, L.R. and D. Chandler, *Theory of the hydrophobic effect.* The Journal of Chemical Physics, 1977. **67**: p. 3683.
141. Alahuhta, P.M., Lunin, V.L. , *Structural and biochemical characterization of C. BESCII CelA.*
142. Guimaraes, B.G., et al., *The crystal structure and catalytic mechanism of cellobiohydrolase CelS, the major enzymatic component of the Clostridium thermocellum cellulosome.* Journal of Molecular Biology, 2002. **320**(3): p. 587-596.
143. Parsiegla, G., et al., *The crystal structure of the processive endocellulase CelF of Clostridium cellulolyticum in complex with a thiooligosaccharide inhibitor at 2.0 angstrom resolution.* Embo Journal, 1998. **17**(19): p. 5551-5562.
144. Parsiegla, G., et al., *Structures of mutants of cellulase Cel48F of Clostridium cellulolyticum in complex with long hemithiocellooligosaccharides give rise to a new view of the substrate pathway during processive action.* J Mol Biol, 2008. **375**(2): p. 499-510.
145. Parsiegla, G., et al., *Crystal structures of the cellulase Cel48F in complex with inhibitors and substrates give insights into its processive action.* Biochemistry, 2000. **39**(37): p. 11238-11246

146. Barr, B.K., et al., *Identification of two functionally different classes of exocellulases.* Biochemistry, 1996. **35**(2): p. 586-592.
147. Irwin, D.C., S. Zhang, and D.B. Wilson, *Cloning, expression and characterization of a Family 48 exocellulase, Cel48A, from Thermobifida fusca.* European Journal of Biochemistry, 2000. **267**(16): p. 4988-4997.
148. Yang, S.J., et al., *Efficient Degradation of Lignocellulosic Plant Biomass, without Pretreatment, by the Thermophilic Anaerobe "Anaerocellum thermophilum" DSM 6725.* Applied and Environmental Microbiology, 2009. **75**(14): p. 4762-4769.
149. ReverbelLeroy, C., et al., *The processive endocellulase CelF, a major component of the Clostridium cellulolyticum cellulosome: Purification and characterization of the recombinant form.* Journal of Bacteriology, 1997. **179**(1): p. 46-52.
150. Olson, D.G., et al., *Deletion of the Cel48S cellulase from Clostridium thermocellum.* Proc Natl Acad Sci U S A, 2010. **107**(41): p. 17727-32.
151. Zhang, Y.H. and L.R. Lynd, *Regulation of cellulase synthesis in batch and continuous cultures of Clostridium thermocellum.* J Bacteriol, 2005. **187**(1): p. 99-106.
152. Caspi, J., et al., *Conversion of Thermobifida fusca free exoglucanases into cellulosomal components: Comparative impact on cellulose-degrading activity.* Journal of biotechnology, 2008. **135**(4): p. 351-357.
153. Rasaiah, J.C., S. Garde, and G. Hummer, *Water in nonpolar confinement: from nanotubes to proteins and beyond.* Annu Rev Phys Chem, 2008. **59**: p. 713-40.
154. Gertz, E.M., *BLAST Scoring Parameters.* 16 March 2005.
155. Guex, N. and M.C. Peitsch, *SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.* Electrophoresis, 1997. **18**(15): p. 2714-23.
156. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* Journal of molecular graphics, 1996. **14**(1): p. 33-38
157. Hong, H., et al., *Role of aromatic side chains in the folding and thermodynamic stability of integral membrane proteins.* J Am Chem Soc, 2007. **129**(26): p. 8320-7.
158. Liu, P., et al., *Observation of a dewetting transition in the collapse of the melittin tetramer.* Nature, 2005. **437**(7055): p. 159-62.
159. Chandler, D., *Interfaces and the driving force of hydrophobic assembly.* Nature, 2005. **437**(7059): p. 640-7.
160. Garde, S., et al., *Origin of Entropy Convergence in Hydrophobic Hydration and Protein Folding.* Phys Rev Lett, 1996. **77**(24): p. 4966-4968.
161. Chelli, R., et al., *Stacking and T-shape competition in aromatic-aromatic amino acid interactions.* Journal of the American Chemical Society, 2002. **124**(21): p. 6133-6143.
162. Boraston, A.B., et al., *Carbohydrate-binding modules: fine-tuning polysaccharide recognition.* Biochem J, 2004. **382**(Pt 3): p. 769-81.
163. Zou, J., et al., *Crystallographic evidence for substrate ring distortion and protein conformational changes during catalysis in cellobiohydrolase Ce16A from trichoderma reesei.* Structure, 1999. **7**(9): p. 1035-45.
164. Becker, D., et al., *Engineering of a glycosidase Family 7 cellobiohydrolase to more alkaline pH optimum: the pH behaviour of Trichoderma reesei CeI7A and its E223S/A224H/L225V/T226A/D262G mutant.* Biochemical Journal, 2001. **356**: p. 19-30.
165. Guimaraes, B.G., et al., *The crystal structure and catalytic mechanism of cellobiohydrolase CelS, the major enzymatic component of the Clostridium thermocellum Cellulosome.* J Mol Biol, 2002. **320**(3): p. 587-96.
166. Payne, C.M., et al., *Multiple functions of aromatic-carbohydrate interactions in a processive cellulase examined with molecular simulation.* J Biol Chem, 2011. **286**(47): p. 41028-35.
167. Wohlert, J., U. Schnupf, and J.W. Brady, *Free energy surfaces for the interaction of D-glucose with planar aromatic groups in aqueous solution.* J Chem Phys, 2010. **133**(15): p. 155103.

168. Parsiegla, G., et al., *Crystal structures of the cellulase Ce148F in complex with inhibitors and substrates give insights into its processive action.* Biochemistry, 2000. **39**(37): p. 11238-11246.
169. Ryckaert, J.-P., G. Ciccotti, and H.J.C. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of< i> n</i>-alkanes.* Journal of Computational Physics, 1977. **23**(3): p. 327-341.
170. Chen, M., et al., *Molecular dynamics simulations of the interaction of glucose with imidazole in aqueous solution.* Carbohydr Res, 2012. **349**: p. 73-7.
171. Himmel, M.E., et al., *Biomass recalcitrance: Engineering plants and enzymes for biofuels production.* Science, 2007. **315**(5813): p. 804-807.
172. DOE., U.S., *Breaking the Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda.* US Dep. Energy, 2006. **Report from the December 2005 Workshop, DOE/SC-0095. U.S. Department of Energy Office of Science**.
173. Alahuhta, P.M., Lunin, V.V. , *Structural and biochemical characterization of C. BESCII CelA.* Submitted (FEB-2012) to the PDB data bank.
174. Kruus, K., et al., *Product inhibition of the recombinant CelS, an exoglucanase component of the Clostridium thermocellum cellulosome.* Applied Microbiology and Biotechnology, 1995. **44**(3-4): p. 399-404.
175. Lamed, R., et al., *Major Characteristics of the Cellulolytic System of Clostridium-Thermocellum Coincide with Those of the Purified Cellulosome.* Enzyme and Microbial Technology, 1985. **7**(1): p. 37-41.
176. Morag, E., et al., *Isolation and Properties of a Major Cellobiohydrolase from the Cellulosome of Clostridium-Thermocellum.* Journal of Bacteriology, 1991. **173**(13): p. 4155-4162.
177. Bu, L., et al., *Probing carbohydrate product expulsion from a processive cellulase with multiple absolute binding free energy methods.* J Biol Chem, 2011. **286**(20): p. 18161-9.
178. Hendrix, D.A. and C. Jarzynski, *A "fast growth" method of computing free energy differences.* Journal of Chemical Physics, 2001. **114**(14): p. 5974-5981.
179. Jarzynski, C., *Nonequilibrium equality for free energy differences.* Physical Review Letters, 1997. **78**(14): p. 2690-2693.
180. Jarzynski, C., *Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach.* Physical Review E, 1997. **56**(5): p. 5018-5035.
181. Efron, B., *The Jackknife, the Bootstrap and Other Resampling Plans.* Society for Applied Mathematics, 1982.
182. Efron, B., Tibishirani, R. J., *An Introduction to the Bootstrap.* 1993.
183. Xiong, H., et al., *Free energy calculations with non-equilibrium methods: applications of the Jarzynski relationship.* Theoretical Chemistry Accounts, 2006. **116**(1-3): p. 338-346.
184. Liphardt, J., et al., *Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski's equality.* Science, 2002. **296**(5574): p. 1832-1835.
185. Bu, L.T., et al., *Product Binding Varies Dramatically between Processive and Nonprocessive Cellulase Enzymes.* Journal of Biological Chemistry, 2012. **287**(29): p. 24807-24813.
186. Strobel, H.J., F.C. Caldwell, and K.A. Dawson, *Carbohydrate Transport by the Anaerobic Thermophile Clostridium thermocellum LQRI.* Appl Environ Microbiol, 1995. **61**(11): p. 4012-5.
187. Kadam, S.K. and A.L. Demain, *Addition of Cloned Beta-Glucosidase Enhances the Degradation of Crystalline Cellulose by the Clostridium-Thermocellum Cellulase Complex.* Biochemical and Biophysical Research Communications, 1989. **161**(2): p. 706-711.
188. Lamed, R., et al., *Efficient Cellulose Solubilization by a Combined Cellulosome-Beta-Glucosidase System.* Applied Biochemistry and Biotechnology, 1991. **27**(2): p. 173-183.

189. Gefen, G., et al., *Enhanced cellulose degradation by targeted integration of a cohesin-fused beta-glucosidase into the Clostridium thermocellum cellulosome.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(26): p. 10298-10303.
190. Frankel, A.E., et al., *Ricin toxin contains at least three galactose-binding sites located in B chain subdomains 1α, 1β, and 2γ.* Biochemistry, 1996. **35**(47): p. 14749-14756
191. Alahuhta, M., et al., *The Unique Binding Mode of Cellulosomal CBM4 from Clostridium thermocellum Cellobiohydrolase A.* Journal of molecular biology, 2010. **402**(2): p. 374-387
192. Himmel, M.E., et al., *Biomass recalcitrance: engineering plants and enzymes for biofuels production.* science, 2007. **315**(5813): p. 804-807
193. Becker, D., et al., *Engineering of a glycosidase Family 7 cellobiohydrolase to more alkaline pH optimum: the pH behaviour of Trichoderma reesei Cel7A and its E223S/A224H/L225V/T226A/D262G mutant.* Biochem. J, 2001. **356**: p. 19-30.
194. Mason, P.E., et al., *Glucose interactions with a model peptide.* Proteins: Structure, Function, and Bioinformatics, 2011. **79**(7): p. 2224-2232
195. Wohlert, J., U. Schnupf, and J.W. Brady, *Free energy surfaces for the interaction of D-glucose with planar aromatic groups in aqueous solution.* The Journal of chemical physics, 2010. **133**: p. 155103.
196. Cuneo, M.J., et al., *The crystal structure of a thermophilic glucose binding protein reveals adaptations that interconvert mono and di-saccharide binding sites.* Journal of molecular biology, 2006. **362**(2): p. 259-270.
197. Alahuhta, M., et al., *The Unique Binding Mode of Cellulosomal CBM4 from Clostridium thermocellum Cellobiohydrolase A.* Journal of molecular biology, 2010. **402**(2): p. 374-387.
198. Vanommeslaeghe, K., et al., *CHARMM general force field: A force field for drug‐like molecules compatible with the CHARMM all‐atom additive biological force fields.* Journal of computational chemistry, 2010. **31**(4): p. 671-690
199. Tavagnacco, L., et al., *Sugar-binding sites on the surface of the carbohydrate-binding module of CBH I from Trichoderma reesei.* Carbohydrate research, 2011. **346**(6): p. 839-846.
200. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water.* The Journal of chemical physics, 1983. **79**: p. 926.
201. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* Journal of molecular graphics, 1996. **14**(1): p. 33-38.
202. Shallenberger, R.S., *Advanced Sugar Chemistry: Principles of Sugar Sterochemistry.* AVI Publshing Company, Westport, Connecticut, 1982.
203. Domanska, U., M.K. Kozlowska, and M. Rogalski, *Solubilities, partition coefficients, density, and surface tension for imidazoles+ octan-1-ol or+ water or+ n-decane.* Journal of Chemical & Engineering Data, 2002. **47**(3): p. 456-466.
204. Liem, S.Y., M.S. Shaik, and P.L. Popelier, *Aqueous imidazole solutions: A structural perspective from simulations with high-rank electrostatic multipole moments.* The Journal of Physical Chemistry B, 2011. **115**(39): p. 11389-11398.
205. Singh, G., et al., *Peptide aggregation in finite systems.* Biophysical journal, 2008. **95**(7): p. 3208-3221.
206. Mason, P.E., et al., *The structure of aqueous guanidinium chloride solutions.* Journal of the American Chemical Society, 2004. **126**(37): p. 11462-11470.
207. Tavagnacco, L., J.W. Brady, and A. Cesàro, *The Interaction of Sorbitol with Caffeine in Aqueous Solution.* Food Biophysics, 2013: p. 1-7.